Understanding Feature Selection in Functional Magnetic Resonance Imaging

Jay Pujara

May 2005

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

Thesis Committee

Tom Mitchell, Chair
Geoff Gordon

Submitted in partial fulfillment of the requirements
for the Degree of Master of Science

Abstract

The advent of functional Magnetic Resonance Imaging (fMRI) has provided researchers with a method of probing the activity of the brain with fine granularity. One goal of fMRI research is to use brain activity to classify the task a subject is performing. Due to the vast quantity of data in fMRI studies, feature selection techniques are used to select relevant brain areas (features) as input to a classifier. In this problem setting, features can be selected through two methods: (1) discriminability tests that compare the feature's activity between the specific tasks to be classified or (2) activity tests that compare the task activity against activity during a rest condition.

A surprising result is that features chosen using activity tests often provide better classification accuracy than discriminability tests, despite receiving no information pertinent to the classification task. The goal of my research has been to use a plausible model of fMRI data coupled with mathematical analysis and experiments using synthetic data to explain this phenomenon. Realistic parameters to this model allow very specific predictions about the performance of different feature selection methods. Subsequently, this approach is validated and extended through the investigation of experimental data in a semantic categories experiment. Finally, the results of this approach are used to explore alternative feature selection methods that have the potential to outperform activity-based feature selection.

A case study comparing feature selection methods in experimental data serves to underscore the applications of this research. Although this research examines fMRI data, the research methodology, analytical approach, and conclusions of this work have the potential to apply to a broad spectrum of problems and provide insight into the general problem of dimensionality reduction.

Executive Summary

The process of selecting relevant inputs for a classifier, known as feature selection, has shown marked benefits in functional imaging experiments that seek to classify the activity of the brain by the task the subject is performing. Three studies described in [Mitchell *et al.*, 2004] show the overwhelming superiority of activity-based feature selection, selection using tests that contrast task activity against fixation (rest) activity, over discriminability-based feature selection, which uses tests that contrast activity between two tasks. Given that the final task is classification, the result that activity-tests outperform discriminability tests is surprising. This paper seeks to answer two fundamental questions about feature selection:

1. Why do activity tests outperform discriminability tests in feature selection?
2. How can the performance of activity-based feature selection be improved through new algorithms?

The approach to answering this questions involves three components: theoretical analysis of a model, tests of algorithms on synthetic data, and experimental validation on an fMRI dataset. The problem of feature selection is specified using a model of fMRI data that splits the regions of the brain (voxels) into three general categories: no-signal voxels, nondiscriminatory signal voxels, and discriminative voxels. No-signal voxels show no activity beyond noise, nondiscriminatory signal voxels cannot differentiate between task conditions but demonstrate activity during tasks, and discriminative voxels can be used to discriminate between tasks.

Theoretical analysis makes an assumption that variances are equal temporally, between experimental conditions, as well as spatially, across all voxels. Using statistical reasoning, this analysis produced a function that expresses the probability of making an error in feature selection in terms of the parameters of the model. These results apply when simple feature selection methods that examined the magnitude of the difference between task activity and fixation in the case of an activity test or between two tasks in the case of a discriminability test, were used to select the most relevant voxel. These probability functions confirm the common expectation that discriminability tests outperform activity tests in feature selection under most circumstances. To better understand the conditions that might result in better selection accuracy from activity tests, a selective exploration of the parameter space was performed.

Results from theoretical analysis showed that a small difference in the mean activity of the tasks, the existence of a single nondiscriminatory signal voxel, or a standard deviation less than twice the mean were all sufficient to allow better feature selection using discriminability-based methods. The result of this analysis is a prediction that superior activity-based performance may be the result of small differences between tasks or very high noise. Additionally, the impact of experimental design was investigated in theoretical analysis, showing that additional task data had a significant impact on the accuracy of feature selection.

Synthetic data experiments endeavored to verify the trends discovered in theoretical analysis and extend the analysis to more complicated feature selection methods. Using the same assumptions and selection algorithms as theoretical analysis confirmed the results of

theoretical analysis, but relaxing some of the assumptions of the previous analysis allowed a critical insight into the operation of feature selection. A more realistic version of feature selection that used variance estimated from experimental data to make selection decisions was introduced. Results of parameter space explorations using these algorithms on synthetic data showed that a small separation in task means was not sufficient to allow accurate feature selection when the variance was estimated empirically. Additionally, an exploration of variance showed that discriminability-based feature selection is profoundly sensitive to noise in the data, and that after a specific threshold performance decreases significantly. These findings suggest an answer to the first question, claiming that noisy data and relatively similar task activity is the cause of poor feature selection using discriminability-based methods.

Synthetic data also suggested an answer to the second question posed by this study, by testing an algorithm referred to as three-way selection. This algorithm creates a model of the training data and uses this model to predict the experimental condition of the data, which includes the fixation condition. The performance of the three-way feature selection algorithm was superior to that of activity-based feature selection in every test. The results of synthetic data analysis predict that three-way feature selection has the ability to outperform activity-based feature selection in functional imaging experiments. Since the results of synthetic data were coupled to assumptions made by the data model, experiments on real fMRI data were used to validate the model parameters as well as the findings of synthetic analysis.

Analysis of experimental data had three goals: (1) validate the assumptions of the model used in the theoretical and synthetic analysis, (2) examine the actual features chosen by each feature selection algorithm and interpret their salient characteristics and (3) compare the performance of feature selection algorithms in a semantic categories experiment. Through a series of histograms, fMRI activity showed many noisy, inactive voxels, a small population of active voxels, and few discriminative voxels; the vast majority of voxels exhibited very low variance, usually smaller than assumed by the model. These results validate the parameters used in the models and lend credence to the predictions made through the analysis of the model.

Next, the top features selected by each algorithm were examined. Activity-based selection consistently chose active voxels with low variance and incidental discriminability. Discriminability-based feature selection consistently chose voxels with large differences between task means which often had very high variance. Three-way feature selection chooses voxels that have low variance and high activity and discriminability. The results of this analysis are that the high variance of discriminability-based analysis is a likely factor behind the poor performance of discriminability-based feature selection, and that three-way selection embodies the desirable qualities of both algorithms: low variance and high discriminability.

The final test of the predictions of the analysis presented in this study is the comparison of feature selection algorithms in experimental data. An initial, primitive analysis using a weak methodology yielded insignificant results, so a more elaborate and computationally-intensive methodology with a greater degree of cross-validation was necessary. The results of this analysis show better performance overall from the three-way selection algorithm, however activity-based feature selection still shows better performance in some subjects. Although three-way selection is not clearly the optimal feature selection algorithm, the desirable qualities of the voxels selected by the algorithm as well as the overall accuracy in experimental data show that the method has great potential.

1: Introduction:

1.1 Motivation

In the last decade, a variety of different functional Magnetic Resonance Imaging (fMRI) experiments have been conducted in hopes of understanding a variety of phenomena that occur in the human brain. While one common method of analysis is to find statistical differences in brain activity during different tasks, a more difficult problem is trying to predict the task based on the brain activity. Machine Learning approaches seek to learn a "cognitive state" [Mitchell *et al.*, 2003] associated with each experimental condition where a task is performed and then use a classifier to discriminate between these cognitive states. In such a studies, the goal is to develop a function of the form *f: fMRI(t,t+n)* $\rightarrow Y$, where *fMRI(t,t+n)* is the data, Y is a discrete set of experimental conditions, and f is a function that uses knowledge about cognitive states to predict a label for the data.

By discriminating between cognitive states, experimenters can differentiate between the different tasks performed in each experimental condition, as well as understand the differing cognitive requirements associated with those tasks. Learning this function from a series of brain data to an experimental condition label requires surmounting many challenges, including the volume of data. Since fMRI experiments produce a great deal of data, on the order of 15,000 readings each second, it is necessary to apply some feature selection method to make learning tractable and prevent overfitting due to spurious correlations.

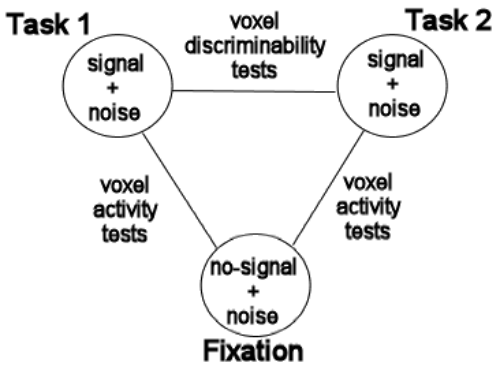

Figure 1: Experimental Paradigm for fMRI Experiments

One interesting facet of fMRI studies that assists in the feature selection process is the experimental paradigm employed when gathering data. Specifically, in fMRI studies, data is collected from a number of experimental conditions in which a specific task is performed, but also includes data from a "fixation" or rest state. This general framework (figure 1) is not specific to fMRI experiments; the circumstances are widespread, including problems such as voice recognition systems that must differentiate between two speakers but also have access to periods of silence, or attempts to detect the user at a computer terminal that also have access to the background processes that run when no user is present, or a system for visually recognizing objects in front of a camera that has access to an empty visual field.

The existence of fixation data allows us to group feature selection methods into two general categories: (1) activity-based feature selection that contrasts activity differences between a fixation state and a task state or (2) discriminability-based feature selection that contrasts two different task states. Given that the desired classification is determining a "cognitive state", i.e. deciding between two task conditions, the discriminability-based feature selection method would naturally be expected to maximize classification accuracy. Contrary to common intuition, feature selection experiments find that activity-based feature selection routinely outperforms discriminability-based feature selection. In fact, [Mitchell *et al.*, 2004] find that activity-based feature selection performs better than discriminability-based feature selection in over 80% of subjects in data collected from three different studies. A result so contrary to common intuition merits further study. Why is activity a better selection metric than discriminability when the task itself is discrimination? What conditions must hold for this result to come about? How can feature selection be improved by leveraging this knowledge?

1.2 Related Work:

The exploration presented here is very directly motivated by [Mitchell *et al.*, 2004] who first commented on this phenomenon which they referred to as the zero-signal learning setting. Specifically, the paper chronicled trends in three vastly different fMRI studies, such as semantic categories, syntactic ambiguity and sentence/picture verification. The remarkable finding was that in 23 of 28 subjects over this assortment of experiments, feature selection that used a simple activity-based method outperformed a discriminability-based method,

while discriminability-based methods only outperformed activity-based methods in one subject. Their approach to activity-based feature selection included three variants involving t-tests between task conditions. The first variant selected active regions regardless of location, the second method set a specific quota for each anatomical subdivision of the brain, and the third method averaged the activation from the selected regions in a given anatomical subdivision. Each of these methods outperformed discriminability-based feature selection in the three experiments studied. The robust pattern of superior results from activity-based feature selection was truly surprising, and an analysis of the results was presented.

[Mitchell *et al.*, 2004] suggested that discriminability-methods are more likely to overfit the data while activity-based methods might select nondiscriminative regions of the brain. They predicted that factors such as a high-dimensional data with many irrelevant features, high-noise conditions, and small training sets would all have an adverse effect on discriminability-based methods. Moreover, analysis presented in their work showed that activity-based methods showed less overfitting and had fewer instances of disjoint distributions that might bias the classification algorithm. One proposal, reinforced by a scatter plot of standard deviations, was that the discriminability tests were susceptible to choosing regions with low signal-to-noise values. One aim of this work is to fully explore the ramifications of the suggestion. The other goal is to approach the problem from a theoretical perspective and attempt to apply theoretical results to experimental data. For example, [Eagle, 2002] attempted to relate feature selection with the number of irrelevant features in a data set, as well as determine the correlation between test error and available training data. Although this work has theoretical contributions, it only considers feature selection that uses discriminability tests and fails to extend the analysis to real data. It is important to note that many papers have compared various feature selection algorithms and the methodology used in applying them in the broad scope of machine learning, but none seem to offer insight to the inconsistency of activity and discriminability methods for feature selection. The cited study is the only research to my knowledge that establishes the questions that arise when the two broad classes of feature selection are analyzed.

2: The Problem Definition:

2.1 Goals of This Research

The goal of this paper, at the broadest level, is to quantify the characteristics that cause activity tests to outperform discriminability tests and apply this knowledge to improve feature selection. The two questions that this paper seeks to answer are

1. Why do activity-based tests outperform discriminability-based tests in feature selection?
2. Is there another feature selection algorithm that can exceed the performance of activity-based feature selection?

This overarching goal requires an approach that tackles the problem in a logical manner and decomposes the more difficult question into smaller, tractable components. The approach adopted to understand feature selection in this study is three-fold:

1. Derive analytical formulas that characterize feature selection
2. Perform experiments on synthetic data to extend and test analytical findings
3. Use knowledge from exploration to interpret the experimental data

Although the primary goal of this work is to understand feature selection, through the path of exploration we hope to encounter facts or ideas that can help improve feature selection. Beyond understanding feature selection, the other major goal of this study is to use an understanding of feature selection to influence and improve the feature selection algorithms used in machine learning. This goal has a practical application, and can be evaluated by using features selected by these algorithms as input to a classifier and examining the resulting accuracy. To achieve the set of goals listed here, the first step is to quantify the real problem and any assumptions that are made to facilitate the analysis of the problem.

2.2 A Description of fMRI Data:

Real data from functional Magnetic Resonance Imaging consists of a series of three-dimensional volumes recorded over an interval of time. Each element of the three-dimensional volume is referred to as a *voxel*, and the data recorded for each voxel at each time step is a decimal value referred to as the hemodynamic response, correlating to the firing rate of a population of neurons. Data may exhibit correlations both spatially, between connected

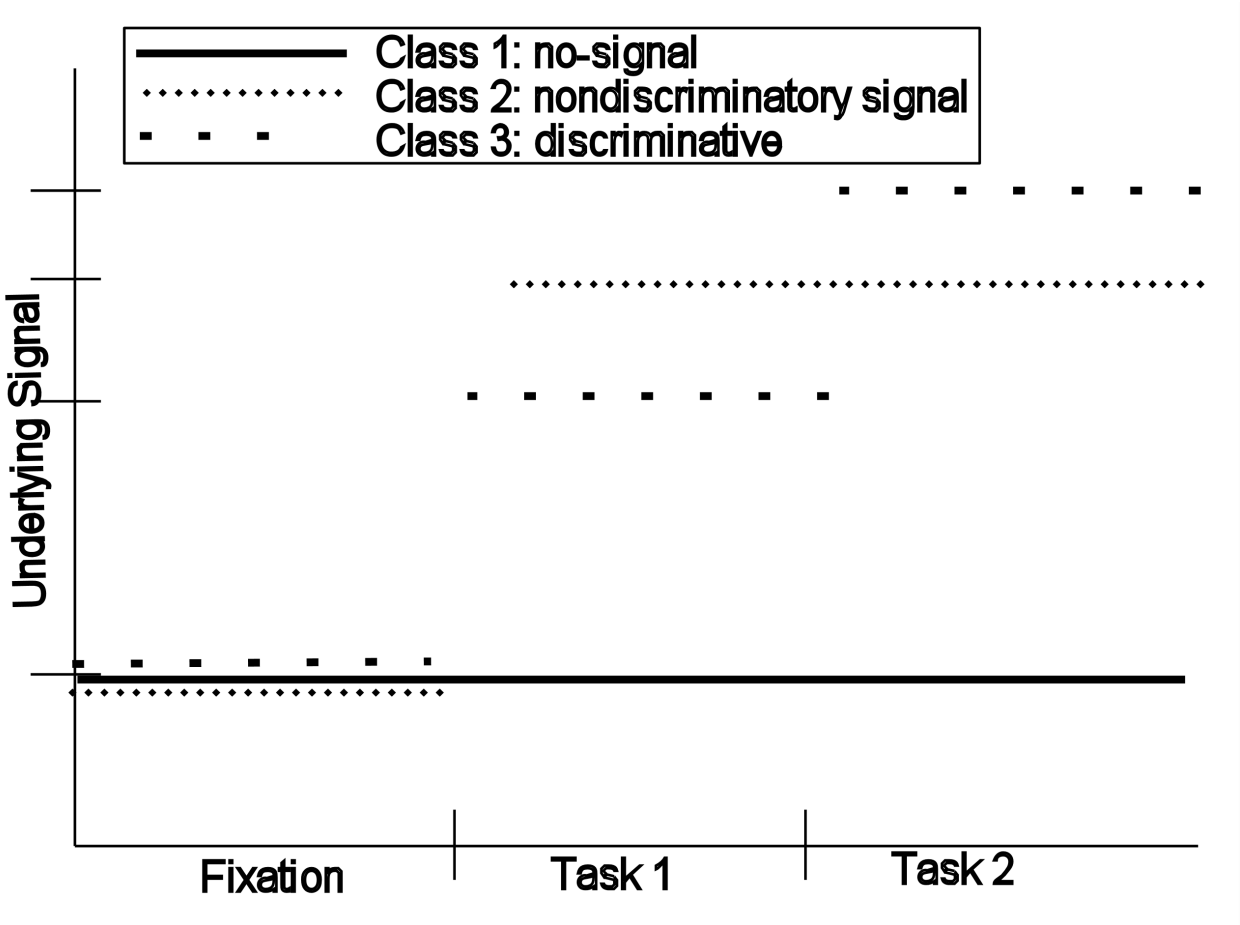

Figure 2: Illustration of three possible voxel classes found in fMRI experiments

regions or neighboring voxels, as well as temporally, following a well-known time course in response to a stimulus. The overall data may be very sparse, with few voxels relevant to a classification task, and data may be noisy as well, causing relevant voxels to exhibit activity that appears uncorrelated to the task during some time steps or facilitating a spurious correlation between task activity and the response of an irrelevant voxel. The true characteristics of fMRI data are difficult to define, and as a result it is useful to consider a model of the data rather than attempting to reason about fMRI data itself.

2.3 A Model of fMRI Data:

The model of fMRI data created for the purposes of this paper strives to maintain realistic properties while providing an opportunity for analytical inquiry. Voxel data in this model is independent both spatially and temporally, and assumed to come from Gaussian normal distributions. Moreover, the experimental design of the hypothetical study is such that

data can come from three different experimental conditions, abbreviated as f, T1, T2 which represent respectively a fixation condition where **no** task activity occurs and two distinct task conditions, Task 1 and Task 2. Beyond an account of the distribution of the data itself, an extremely simplified model of the neuron populations found in the brain is used to elucidate trends. To model the behavior of these voxels, we assume that each voxel can belong to one of three distinct classes. The first class of voxels have no signal and exhibit truly random activity indistinguishable from noise, the data retains the same distribution between a fixation condition and task conditions. The second class of voxels display a signal during the task conditions, but this signal is not unique between task conditions and cannot be used to discriminate between two tasks. The final class of voxels considered in this problem statement are those that are truly discriminative, that is to say those voxels whose distribution is differentiable based on the task conditions. An illustration of these three classes is shown in Figure 2.

The three classes we describe can be formalized as follows. We define E to be the experimental condition, such that $E \in \{f, T1, T2\}$, where f is fixation, T1 is Task 1, T2 is Task 2. Then, for some voxel V drawn from the data set of all voxels in the entire brain or in some specific region, The distribution of this voxel's activity can be expressed as conditionally dependent on both the experimental condition as well as which of the three populations of voxel activity it belongs to as shown in Table 1.

Class 1 (no-signal)	$\forall V \in C1 : P(V = x \mid E = f) =$ $P(V = x \mid E = T1 \vee T2) = P(V = x) \sim N(0, \sigma_f^2)$
Class 2 (nondiscriminatory signal)	$\forall V \in C2 : P(V = x \mid E = f) \sim N(0, \sigma_f^2)$ $P(V = x \mid E = T1 \vee T2) = N(\mu_A, \sigma_a^2)$
Class 3 (discriminative)	$\forall V \in C3 : P(V = x \mid E = f) \sim N(0, \sigma_f^2)$ $P(V = x \mid E = T1) \sim N(\mu_{T1}, \sigma_{t1}^2)$ $P(V = x \mid E = T2) \sim N(\mu_{T2}, \sigma_{t2}^2)$

Table 1: Distribution of voxels by class

2.4 Problem Setting:

The ultimate goal of this model is to capture the important trends in different feature selection strategies and lend insight into the conditions that comprise actual fMRI data. If the data is characterized by this model, a number of parameters establish the problem space

and can be classified into parameters that are established by the data as well as parameters that result from the experimental design or feature selection process. These parameters, referred to as Ψ, are shown below in Table 2.

n_{C1}:	Number of Class 1 voxels
n_{C2}:	Number of Class 2 voxels
n_{C3}:	Number of Class 3 voxels
σ_f:	Variance of data for all voxels in the fixation condition
μ_A:	Mean value of Class 2 voxels during task activity
σ_A:	Variance of Class 2 voxels during task activity
μ_{T1}:	Mean value of Class 3 voxels during Task 1
σ_{T1}:	Variance of Class 3 voxels during Task 1
μ_{T2}:	Mean value of Class 3 voxels during Task 2
σ_{T2}:	Variance of Class 3 voxels during Task 2
n_f:	Number of trials in the fixation condition
n_{T1}:	Number of trials in the Task 1 condition
n_{T2}:	Number of trials in the Task 2 condition
n_V:	Number of voxels chosen by our feature selection algorithm
Γ:	Feature selection algorithm used to choose relevant voxels

Table 2: Complete Model Parameters: Ψ

Our goal is to derive some function F to map our parameters, Ψ, to the expected error, ϵ, of the feature selection process, where error is defined as the probability of choosing irrelevant voxels during selection, $F(\Psi) = \epsilon$. The parameter space is quite large, and only a few of the parameters, namely those in the lower division of the table, can be controlled by experimental design or the selection of a feature selection algorithm. In order to simplify analysis, a number of assumptions are made about these parameters in the presented analysis of the model. The most significant assumption about the data itself is that all voxels are assumed to have equal variance, $\sigma_f = \sigma_A = \sigma_{T1} = \sigma_{T2} = \sigma$ regardless of the class or experimental condition. Not only does this assumption simplify the analysis and allow the discovery of general trends, but the assumption is also realistic based on profiles of fMRI data. The remaining assumptions and simplifications involve the heart of the problem, the feature selection method, Γ.

2.5 Feature Selection Algorithms

Many different feature selection algorithms have been used in fMRI studies, including t-tests [Mitchell *et al.*, 2004], analysis of variance (ANOVA) [Cox and Savoy, 2003], correlation to a simulated hemodynamic response, [Cox and Savoy, 2003], linear regression [Friston *et al.*, 1995]. One of the most popular feature selection methods is the t-test, which attempts to determine how significantly two random variables differ using the t-statistic, $t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$. The t-test will be used to help formulate our model of feature selection.

As assumptions have been made to preserve the important characteristics of fMRI data without the accompanying complexity, similar assumptions are made to preserve the performance of feature selection algorithms without delving into the complexity and variety of different incarnations of these algorithms. By defining the data in terms of normal distributions and making assumptions about variance, much of the groundwork for this approach has already been laid; what remains is to concisely define and distinguish activity and discriminability tests. The assumption made in the previous section, that variances should be considered equal between conditions and classes, greatly simplifies this task.

If the feature selection algorithm also assumes that variances are equal, an activity test merely needs to select the most active voxel and a discriminability test needs to select the voxel with the greatest difference in means. The most significant ambiguity in this description is the exact comparison made in the activity test: how should comparisons be made between task conditions and fixation? To simplify the analysis even further, we assume an activity test compares the mean activity of the Task 1 condition to fixation. Based on this discussion, a formal description of the feature selection algorithm, Γ which receives as input the dataset, D, is given as:

$$\begin{aligned} \Gamma(D) &= max_V(\mu_{T1} - \mu_f) && \text{Activity} \\ \Gamma(D) &= max_V(|\mu_{T1} - \mu_{T2}|) && \text{Discriminability} \end{aligned}$$

where max_V represents choosing the top n_V voxels in the population that have the largest values for the appropriate contrast. Since the variance of the data are assumed to be equal, the most active voxel is the one whose difference from the mean is the highest. Similarly, the most discriminative voxel is the one whose difference in means is the highest. Allowing these simplifications creates very simple formulas for both feature selection methods that facilitate

an intuitive understanding of what occurs in feature selection. Despite these simplifications, the derivation of F is still complex and often general trends will have to suffice for an actual mathematical relation between these quantities.

2.6 Remarks

Throughout the discussion of the model, we have noted various assumptions and have attempted to defend their legitimacy. A collection of these assumptions, a review of their legitimacy and some perspective on the divergence from real fMRI data can provide a useful disclaimer to the analysis presented here. The most important principle that is violated through the creation of a model is the continuum of different parameters in fMRI data. For example, creating three discrete classes of voxels is an extreme oversimplification of the real state of affairs where each voxel has a different mean, variance, and task-discriminability. Another assumption that lies at the core of this analysis is that fMRI data can be modeled using a normal distribution. While the actual hemodynamic response of the brain might have a different distribution, the assumption of normally distributed data is made by the Gaussian Naive Bayes classifier with marked success in many fMRI experiments.

It is important to note that creating any model of fMRI data is apt to introduce assumptions which will deviate from the true nature of the data. One could imagine a far more elaborate model that would choose the task means and variances for each voxel from some known distribution, creating a diverse set of voxels, and perhaps the insights gained from this model would be more meaningful, if at the cost of a more difficult analysis. However, the true nature of fMRI data is difficult to ascertain and the more parameters necessary to define the model, the greater the potential of making an erroneous assumption about the model. The main advantage of such a discrete and deterministic method of defining the activity of the brain is the overwhelming simplicity it affords. With a more comprehensible mathematical analysis and simply generated synthetic data, this study seeks to penetrate into the underlying nature of fMRI data and feature selection as much as possible through modeling and leave the remainder of the analysis to experiments with real data.

3: Analytical Derivations

3.1 Approach:

Given this model of data, one approach to understanding how activity tests outperform discriminability tests is the use of mathematical analysis of the model itself to provide a relation between the model parameters and the probability of erroneous feature selection. Due to the complexity of the analysis, the results in the following text are restricted to the scenario where only one feature is selected, and only one relevant feature exists in the data set. The goal in this section will be to first derive the formula necessary to allow analysis of feature selection, and then interpret these formula to understand why activity tests outperform discriminability tests. As extracting trends from the derivations alone can be difficult, numerical methods are used to evaluate the functions for different parameter values to show how characteristics of the data affect feature selection.

3.2 A Simple Example:

One way to gain an intuition about the behavior that manifests itself is with a simple scenario where only two voxels - one from Class 1 (no-signal) and one from Class 3 (discriminative) - are considered. Note again that for the sake of simplicity, we assume all variances are equal. In the previous section, the behavior of each class of voxels was described in terms of some true distribution and parameters that affect the distribution. However, the feature selection process will rely on parameter estimates based on the true value of the parameters and the amount of data that is sampled. For this reason, consider three variables n_f, n_{T1}, n_{T2} corresponding to the number of samples from the conditions fixation, Task 1 and Task 2 respectively. What are the expected distributions of the parameters when constrained by the amount of data collected in an experiment? The answer is shown in Table 3; for completeness and future reference, derivations for Class 2 (nondiscriminatory signal) are also included, although our simple example does not include any Class 2 voxels and the immediate discussion will not consider them.

Based on this information, it is necessary to model what happens when activity-based or discriminability-based feature selection is performed. The pure feature selection algorithms

$\hat{\mu}_f$	$N(0, \frac{\sigma^2}{n_f})$
$\hat{\mu}_{T1}$	$N(0, \frac{\sigma^2}{n_1}), V \in C1$ $N(\mu_A, \frac{\sigma^2}{n_1}), V \in C2$ $N(\mu_{T1}, \frac{\sigma^2}{n_1}), V \in C2$
$\hat{\mu}_{T2}$	$N(0, \frac{\sigma^2}{n_2}), V \in C1$ $N(\mu_A, \frac{\sigma^2}{n_2}), V \in C3$ $N(\mu_{T2}, \frac{\sigma^2}{n_2}), V \in C3$

Table 3: Distribution of Parameter Estimates based on Data

described previously allowed an intuitive understanding because the feature selection process was simply finding a maximal value. When the individual contrasts required for feature selection have differing variances, analysis must reason about the *distribution* of the differences in mean values rather than the difference in the mean values alone. In activity-based feature selection, the variable of interest is the difference between the estimated mean of task activity and the estimated mean of fixation activity, or $\hat{\mu}_{T1} - \hat{\mu}_f$, henceforth referred to as α. In discriminability based feature selection, the variable of interest is the absolute value of the difference between the estimated mean of Task 1 and the estimated mean of Task 2, $|\hat{\mu}_{T1} - \hat{\mu}_{T2}|$, henceforth referred to as β. The distribution of these parameters is defined in Table 4

	$\alpha = \hat{\mu}_{T1} - \hat{\mu}_f$	$\beta = \lvert\hat{\mu}_{T1} - \hat{\mu}_{T2}\rvert$
Class 1 (no-signal)	$N(0, \frac{(n_f+n_1)\sigma^2}{n_f * n_1})$	$N(0, \frac{(n_1+n_2)\sigma^2}{n_1 * n_2})$
Class 2 (nondiscriminatory signal)	$N(\mu_A, \frac{(n_f+n_1)\sigma^2}{n_f * n_1})$	$N(0, \frac{(n_1+n_2)\sigma^2}{n_1 * n_2})$
Class 3 (discriminative)	$N(\mu_{T1}, \frac{(n_f+n_1)\sigma^2}{n_f * n_1})$	$[N(\mu_{T1} - \mu_{T2}, \frac{(n_1+n_2)\sigma^2}{n_1 * n_2})$ $+N(\mu_{T2} - \mu_{T1}, \frac{(n_1+n_2)\sigma^2}{n_1 * n_2})]$

Table 4: Distribution of Differences in Parameter Estimates by Class

The crucial problem in this example is to determine the probability of making an error in feature selection, the probability that the irrelevant, Class 1 voxel will be chosen by the feature selection algorithm instead of the the relevant Class 3 voxel. These formulas alone provide some crucial intuitions. When comparing voxels for feature selection, the real comparison is between data sampled from α and β. Notice that the two differences between activity and discriminability tests are the variance of the distributions (henceforth referred to as σ_α^2 and σ_β^2), and in the case of discriminative Class 3 voxels, the mean value of

the distribution. These two factors are responsible for determining which feature selection method will work best. A lower variance for the data will result in more tightly clustered data, less spread, and greater discriminability between the two voxel classes, so a low variance is a desirable trait. Note, additionally, that the number of trials from each condition is equally essential to determining the variance of α and β. The other quantity to consider is the difference in the means - if the task means are very different then they will be easily discriminable from the noise of the data.

These trends will be more apparent if the actual probability of selecting an irrelevant voxel is derived for each of these cases. What is the probability that the no-signal Class 1 voxel is chosen as relevant instead of the discriminatory Class 3 voxel? This question must be answered separately for the two feature selection methods - activity and discriminability - but the approach remains the same for both. The idea is to reason about random variables drawn from the distributions α (distribution of $\hat{\mu}_{T1} - \hat{\mu}_f$) and β (distribution of $|\hat{\mu}_{T1} - \hat{\mu}_{T2}|$) that were introduced earlier and computed in Table 4. The value of the difference between the estimated mean of task 1 and the estimated mean of fixation is a random variable, and conclusions about its value must come in the form of statements of probability. Simply put, we cannot authoritatively state what the maximum value of $\hat{\mu}_{T1} - \hat{\mu}_f$ will be, but we can generate the probability that this contrast is equal to a specific value. In the analysis that follows we use this property to derive the probability that a single Class 1 voxel is judged more relevant than a single Class 3 voxel, and use shorthand to simplify the derivation. V_{C1} is a random variable distributed according to α or β for a Class 1 voxel, as stated in Table 4. Notation such as $\{V_{C1}\}$ is used to represent a set of such random variables. Similarly, V_{C3} is a random variable whose distribution conforms to that of a Class 3 voxel. The specific distribution, α or β is not essential until the final step of the analysis when the probabilities corresponding to the different feature selection methods are substituted into the equation. Note, z is the standard normal (N(0,1)) PDF, and Z is the standard normal CDF

$$
\begin{aligned}
P(\text{ Class 1 voxel appears more relevant than Class 3 voxel}) &= \\
P(\text{ random variable } V_{C1} \text{ has value greater than random variable } V_{C3}) &= \\
P(V_{C1} > V_{C3}) &= \\
\forall\, \theta, P((V_{C1} > \theta) \cap V_{C3} = \theta) &=
\end{aligned}
$$

$$\int_\theta [P(V_{C1} > \theta) * P(V_{C3} = \theta)] \quad =$$

$$\begin{cases} \int_\theta \left[(1 - Z(\frac{\theta}{\sigma_\alpha})) * z(\frac{\theta - \mu_{T1}}{\sigma_\alpha})\right] & \text{(Activity Test)} \\ \int_\theta \left[(1 - Z(\frac{\theta}{\sigma_\beta})) * \left(z(\frac{\theta - \mu_{T1} + \mu_{T2}}{\sigma_\beta}) + z(\frac{\theta - \mu_{T2} + \mu_{T1}}{\sigma_\beta})\right)\right] & \text{(Discrim Test)} \end{cases}$$

Additionally, this derivation can be extended to the case where there are many (n_{C1}) Class 1 voxels and only one Class 3 voxel.

$$\begin{aligned} P(\text{some } \{V_{C1}\} \text{ is more relevant than } V_{C3}) &= \\ P(\text{ not all } \{V_{C1}\} \text{ less relevant than } V_{C3}) &= \\ \int_\theta [(1 - (P(V_{C1} < \theta)^{n_{C1}}) * P(V_{C3} = \theta)] &= \end{aligned}$$

$$\begin{cases} \int_\theta \left[(1 - Z(\frac{\theta}{\sigma_\alpha})^{n_{C1}}) * z(\frac{\theta - \mu_{T1}}{\sigma_\alpha})\right] & \text{(Activity Test)} \\ \int_\theta \left[(1 - Z(\frac{\theta}{\sigma_\beta})^{n_{C1}}) * \left(z(\frac{\theta - \mu_{T1} + \mu_{T2}}{\sigma_\beta}) + z(\frac{\theta - \mu_{T2} + \mu_{T1}}{\sigma_\beta})\right)\right] & \text{(Discrim. Test)} \end{cases}$$

Although this derivation provides a mathematical formula to determine the probability of mistakenly choosing an irrelevant voxel, it relies on the CDF of a normal distribution and so analysis fails to provide the formula in a closed form. Investigating all trends in this example is beyond the scope of this document, but some trends are readily apparent from this expression. For example, as the number of Class 1 voxels (n_{C1}) increases, the expression $Z(\frac{\theta}{\sigma})^{n_{C1}}$ approaches 0, causing the entire expression $(1 - Z(\frac{\theta}{\sigma})^{n_{C1}})$ to approach 1, leaving the integral over all values of V_{C3}, which is simply 1. That is to say, as the number of no-signal voxel increases the probability of mistakenly choosing an irrelevant voxel approaches 1. However, this does not provide a clear idea of how the function approaches the asymptote. This trend is dependent on the variance and means of the functions in question, but numerical integration provides a general form for the function. Specifically, the probability of choosing a Class 1 voxel instead of a Class 3 voxel as a function of Class 1 voxels initially grows linearly and then logarithmically approaches an asymptote. The point at which linear growth gives way to asymptotic growth depends on the means (μ_{T1} and μ_{T2}) as well as the variance of the data. Figure 3 demonstrates this trend in two different examples where parameters vary.

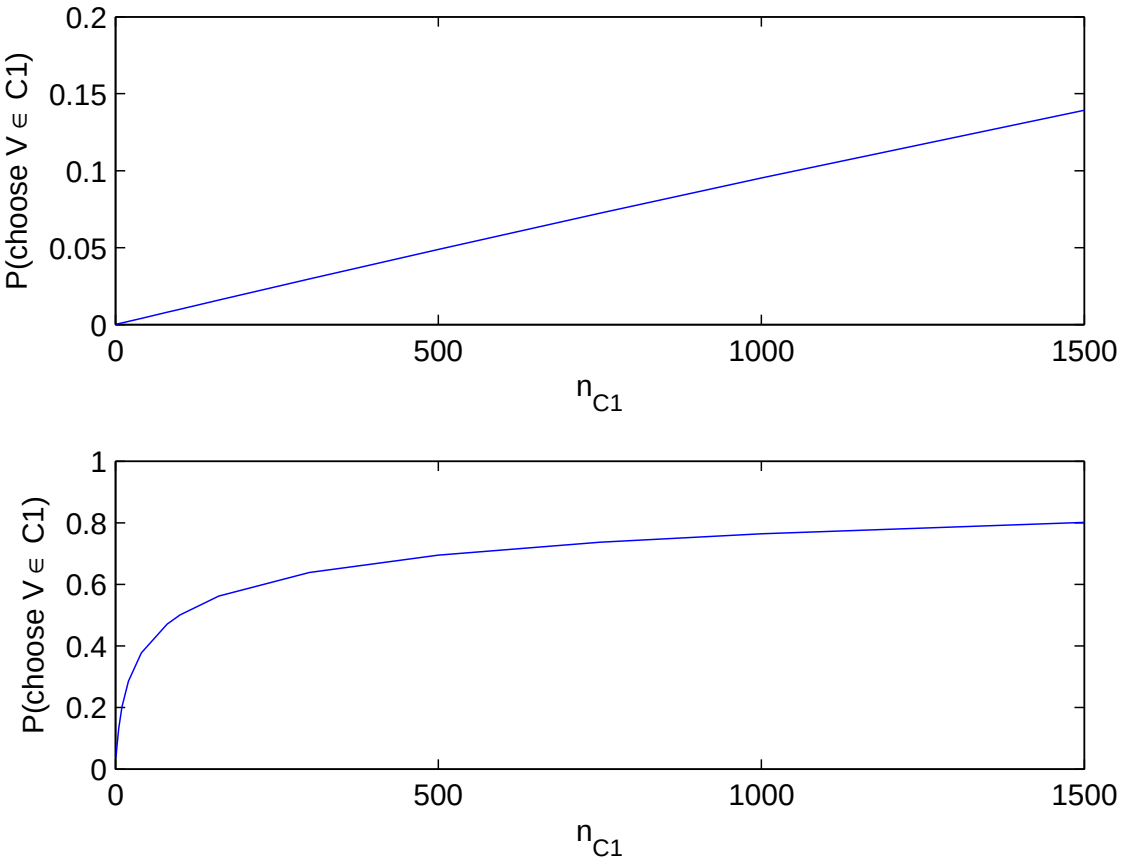

Figure 3: Growth of the probability of choosing an irrelevant voxel under different conditions

3.3 A More General Model

Now that the process of deriving the probability of error for a feature selection algorithm and a given set of voxels is clear, a more complicated derivation can be introduced. If the scenario in the last section is expanded to include a single Class 3 (discriminative) voxel and an arbitrary number of Class 2 (nondiscriminatory signal) and Class 1 (no-signal) voxels, a deeper understanding of the phenomena in question can be obtained. What is the probability that a Class 1 or Class 2 voxel will be judged more relevant than a Class 3 voxel?

These equations yield some interesting observations. When using discriminability tests, the addition of Class 2, nondiscriminatory signal voxels, has the same effect to adding more Class 1, no-signal voxels. This occurs because for both classes of voxels, the expected difference in means is 0 (see Table 4). On the other hand, because the contrast between Task 1 and fixation differs depending on the class of the voxel, the addition of Class 2 voxels has a

$$
\begin{array}{l}
\mathrm{P}(\{V_{C1}\} \text{ more relevant than } V_{C3} \text{ or } \{V_{C2}\} \text{ more relevant than } V_{C3}) = \\
P((\{V_{C1}\} > V_{C3}) \cup (\{V_{C2}\} > V_{C3}) = \\
P(\{V_{C1}\} > V_{C3}) + P(\{V_{C2}\} > V_{C3}) - P((\{V_{C1}\} > V_{C3}) \cap (\{V_{C2}\} > V_{C3}) = \\
\int_\theta P(V_{C3} = \theta) * [(1 - (P(V_{C1} < \theta)^{n_{C1}}) + (1 - (P(V_{C2} < \theta)^{n_{C2}}) \\
\qquad -[(1 - (P(V_{C1} < \theta)^{n_{C1}}) * (1 - (P(V_{C2} < \theta)^{n_{C2}})]] = \\
\begin{cases}
\int_\theta z(\frac{\theta - \mu_{T1}}{\sigma_\alpha}) * [(1 - Z(\frac{\theta}{\sigma_\alpha})^{n_{C1}}) + (1 - Z(\frac{\theta - \mu_A}{\sigma_\alpha})^{n_{C2}}) \\
\qquad -[(1 - Z(\frac{\theta - \mu_A}{\sigma_\alpha})^{n_{C2}}) * (1 - Z(\frac{\theta}{\sigma_\alpha})^{n_{C1}})]] & \text{(Activity Test)} \\
\int_\theta \left(z(\frac{\theta - \mu_{T1} + \mu_{T2}}{\sigma_\beta}) + z(\frac{\theta - \mu_{T2} + \mu_{T1}}{\sigma_\beta})\right) * [(1 - Z(\frac{\theta}{\sigma_\beta})^{(n_{C1} + n_{C2})})] & \text{(Discrim. Test)}
\end{cases}
\end{array}
$$

different impact than Class 1 voxels on an activity test. The net result is that an activity test may be crippled by the addition of Class 2 voxels, as the increased nondiscriminative activity of the Class 2 voxels during task conditions is hard to distinguish from the discriminative activity of Class 3 voxels. However, the surprising result that initiated this work was that activity tests do outperform discriminability tests. How can this observation be reconciled with the equations above?

Consider a few possibilities that might cause activity tests to outperform discriminability test: the discriminative task means (μ_{T1}, μ_{T2}) might be very close together, the mean activity of Class 2 voxels (μ_A) might be very small, the variance of the data may be very large compared with the difference in task means, or there may be very few Class 2 voxels. After investing the labor to generate analytical expressions for the performance of activity and discriminability tests in feature selection, the new goal is to define the boundary conditions where discriminability tests begin to outperform activity tests. The previous hypotheses that posit conditions that might allow activity tests to outperform discriminability tests offer a natural set of boundaries to investigate. Each of these hypotheses can be qualified or disqualified by numerical integration of the equations presented earlier.

3.4 Numerical Integration: Boundary Conditions

n_f	n_{T1}	n_{T2}	n_{C1}	n_{C2}	n_{C3}	$\sigma_{f,T1,T2,A}$	μ_f	μ_A	μ_{T1}	μ_{T2}
300	500	500	10000	10	1	3	0	2.1	2.6	1.5

Table 5: default parameters for this section

The derivations in the previous sections offered mathematical insight to issues that are relevant to the feature selection process. The derivation so far has produced a partial so-

lution mapping the parameters of the data model Ψ (as stated in Table 2) to an expected error ϵ during feature selection, giving us the desired function of form $F(\Psi) = \epsilon$. Yet these intuitions do not necessarily confirm the phenomenon under investigation, namely the superiority of activity tests to discriminability tests. Worse, the immediate intuition from the last derivation contradicts empirical results insofar as suggesting that discriminability tests should perform relatively better as the number of Class 2, nondiscriminatory signal voxels increases. How can this be? One possibility is that a specific set of model parameters might elicit performance similar to that seen in experiments. Four possibilities mentioned in the previous section will be investigated to see what sort of performance results from some specific conditions. The goal of this analysis is to find a set of conditions that might show how activity results outperform discriminability results. Some reasonable default parameters are used in the following experiments unless noted otherwise and are listed in Table 5 for completeness.

- **differences in task means:**

  Assume, in the simplest case, that no Class 2 voxels exist. How does the difference between task means affect the probability of choosing a nondiscriminative voxel during feature selection? Surprisingly, the difference in task means can be small relative to the standard deviation without affecting the probability of erroneous feature selection. For example, with $\mu_{T2} = 1.5$, the probability of choosing an irrelevant voxel, ϵ when $\mu_{T1} = 2.3$ is $\epsilon = .362$ for a discriminability test, and this probability converges to approximately zero once the first task mean becomes larger than 2.6. ($\epsilon \approx 0$ for values of $\mu_{T1} > 2.6$). The probability of an incorrect voxel selection using an activity test remains approximately zero as well throughout the range. Although activity tests outperform discriminability tests in this situation, the key result is that the difference between them becomes insignificant as the difference in task means increases past a certain value. This single example is representative of a general trend that causes accuracy to approach an asymptote as the difference in task means increases beyond some threshold. While the magnitude of this difference is large in absolute terms, it amounts to $.37\sigma$, well under half a standard deviation. The net result of this trend reveals that relatively good feature selection performance is possible on data whose differences in means is small relative to the standard deviation, or that even small signal-to-noise ratio will result in successful classification.

- **number of nondiscriminatory signal voxels:**

  If the distribution of discriminatory voxels is not to blame for the observed differences between activity and discriminability-based feature selection, perhaps the culprit is the number of Class 2, nondiscriminatory signal voxels. Consider the case where the parameters are the same as those listed in Table 5, notably $\mu_{T1} = 2.6, \mu_{T2} = 1.5$, $\mu_A = 2.1$. How many Class 2 voxels would be necessary before discriminability-based feature selection outperformed activity-based feature selection when attempting to select a single discriminative voxel? Surprisingly, even a single Class 2 voxel will result in a .05 error rate for activity-based feature selection and an error rate of .03 for discriminability based feature selection. Moreover, if 55 Class 2 voxels are found in the population, the error rate for selection, $\epsilon_{activity}$ increases to .50, while the error rate of discriminability-based feature selection remains at .03. Even if few nondiscriminatory voxels are in our population, the performance of activity-based feature selection is injured. As such, either the model of data fails to capture the true nature of the problem or the hypothesis that the performance characteristics of feature selection methods is the result of a small number of nondiscriminatory signal voxels is invalid.

- **mean value of nondiscriminatory signal voxel:**

  How active can a Class 2 voxel be before its presence causes activity-test performance to deteriorate? Consider a situation where Class 2 voxels have mean $\mu_A = 1.45$, while the single discriminatory voxel has the default means $\mu_{T1} = 2.6, \mu_{T2} = 1.5$. In this case, Class 2 voxels whose mean value is under the lower activation threshold of a discriminative voxel can have a marked effect on the reliability of feature selection if present in large quantities. With 1000 such voxels, 10% of the population size, the probability of selecting an irrelevant voxel using a discriminability-based method is equal to that of an activity-based method at .03. Also note that this estimate is a best-case scenario for a discriminative voxel during an activity-test; the analysis compares the higher task mean of a discriminative voxel against fixation, yielding a higher probability of successfully choosing that voxel. The result that even relatively inactive nondiscriminatory voxels can have a large impact on the performance of activity-based feature selection, provided there are enough such voxels, can be construed as a general statement about feature selection or a failure of this model. The success of activity-based

feature selection might result from a dataset that truly contains few and relatively inactive nondiscrimantory signal voxels. However, note that the difference between the expected error rates of activity and discriminability feature selections is not very large, a result that belies actual experimental findings.

- **signal variance:**

  How does noise in the voxels affect both activity and discriminability tests? One assumption, albeit reasonable, that has been propagated through the experiments detailed so far is that the data is noisy. Estimates of noise have been conservative, as examples so far have assumed that the mean of task 2 activity, μ_{T2}, was 1.5 while the standard deviation, σ was 3, creating conditions where the magnitude of the noise is substantial compared to the signal in the data. Consider the case where the standard deviation is approximately three times the lower mean task value, $\sigma_{f,T1,T2} = 4.5$. Note that this analysis assumes that there is one discriminatory voxel with means $\mu_{T1} = 2.6, \mu_{T2} = 1.5$ and ten nondiscriminatory signal voxels with mean $\mu_A = 2.1$. In this situation, activity-based feature selection does outperform discriminability-based feature selection; the error rate for the activity selection is .50 while the error rate for the discriminability based feature selection is .49. If noisier data is considered, $\sigma_{f,T1,T2} = 7$, the error rates increase to .68 for activity-based selection and .91 for discriminability-based feature selection. The actual curve, as shown in Figure 4 is quite interested. As noted before, both values for the noise are large with respect to the mean: a possibly unrealistic assumption. However, very noisy data certainly could be a factor in the observed superiority of activity-based feature selection.

These and other trends are apparent in Figure 4, which shows the decay in performance as a single parameter varies from the defaults in Table 5. Additionally, earlier discussions underscored the importance of the amount of data available in an experiment. Controlling the experimental design to collect differing amounts of experimental design is possibly the *only* way a researcher can improve feature selection apart from the selection algorithm. Figure 5 shows the effects of experimental design, demonstrating the effects of varying the amount of fixation data and task data collected while maintaining default values for other parameters. Increasing the amount of data collected serves to reduce the variance in parameter estimates, allowing better feature selection. As the amount of fixation data increases, ac-

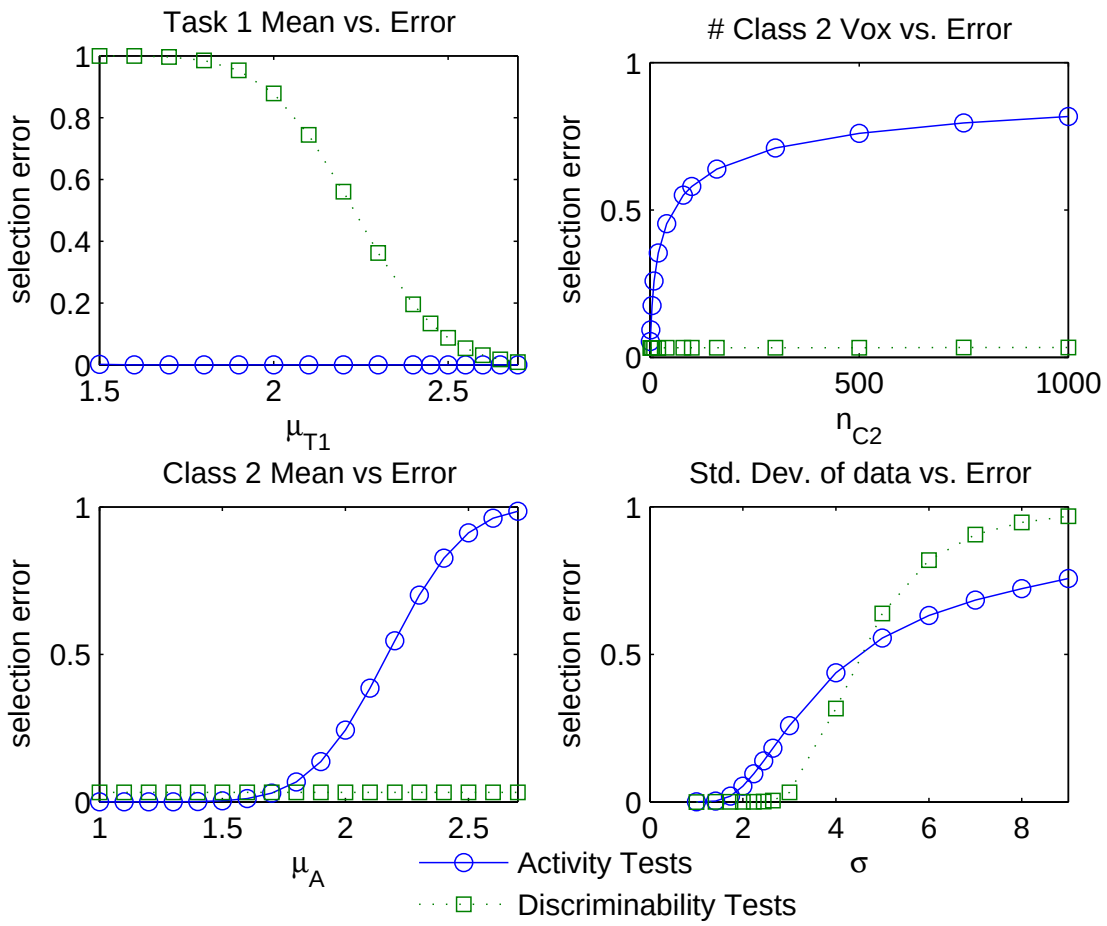

Figure 4: Trends from Numerical Integration

tivity tests have better parameter estimates and show better performance, while increasing the amount of task data collected improves both activity and discriminability-based feature selection. While collecting more data leads to improved results, the practical considerations of patient attentiveness and experimental costs may prevent copious data selection.

The net result of these explorations has been to characterize the performance of feature selection algorithms in a series of constrained situations. While the results have been valuable and contributed to the comprehension of the intricacies of feature selection, the power of the analysis has been limited. The most limiting constraint in the analysis so far has been the ability to model the performance with only one Class 3 discriminative voxel. In practice, there is more than one discriminative voxel, and feature selection methods select more than one voxel as input to feature selection methods. While it might be possible to derive similar formula for multiple discriminative voxels or a selection set larger than one

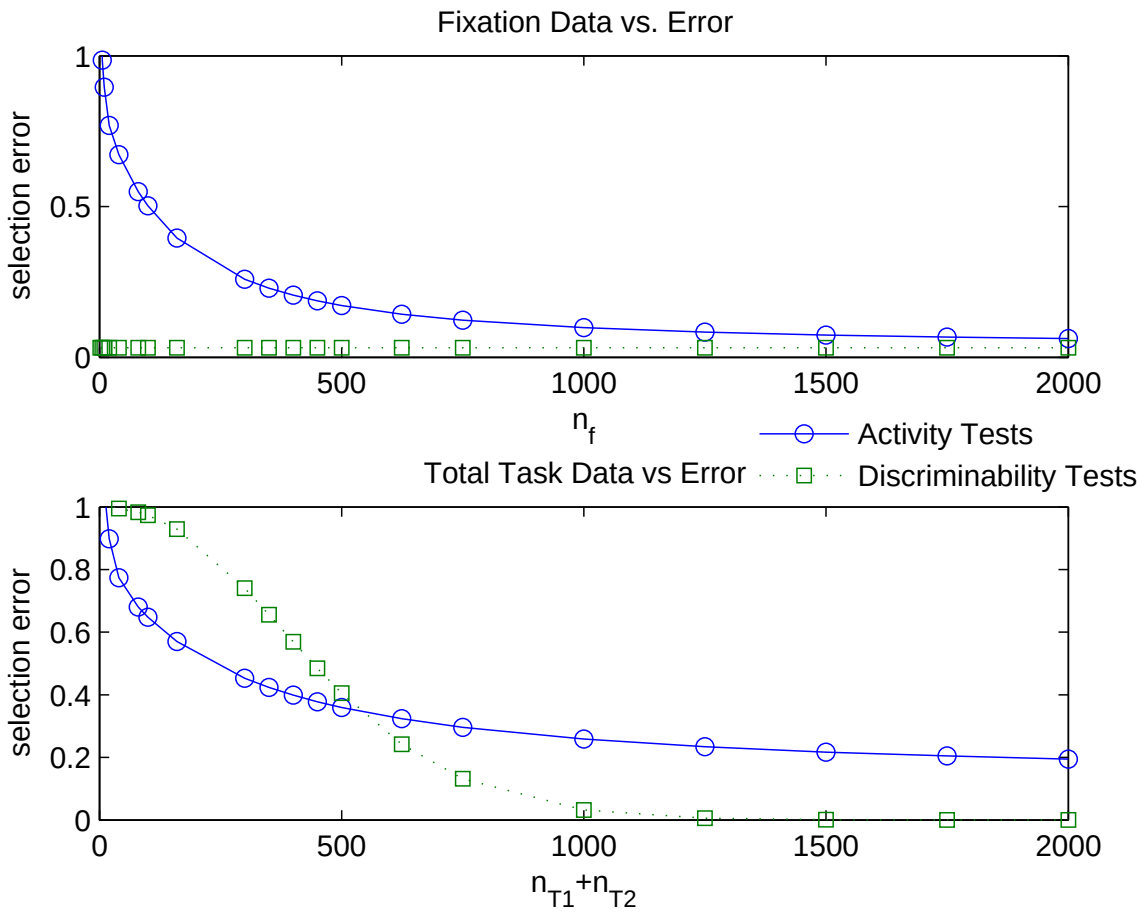

Figure 5: Impact of Experimental Design

voxel, or both, the task seems daunting and was beyond the scope of my knowledge. Instead, to understand more complicated facets of the feature selection problem this study shifted its focus to synthetic data. Moreover, having both synthetic data and theoretical analysis allows a valuable comparison between the theoretical predictions based on a model and the experimental findings from data that is derived from the model parameters.

4: Experiments Using Synthetic Data

4.1 Approach

The analysis of feature selection using synthetic data has many possible avenues, so a coherent approach is necessary. In order to validate the findings of the theoretical approach, synthetic data should produce the same results. Moreover, an additional flexibility allowed

by synthetic data is the ability to define new feature selection algorithms through implementation rather than mathematical analysis. This allows us to test more complicated or involved selection strategies without requiring complex mathematical calculations. These motivations allow us to define the goals of synthetic data analysis.

1. Verify the predictions from theoretical analysis under identical conditions
2. Introduce new feature selection algorithms
3. Apply feature selection algorithms to the synthetic data to extend the predictions of theoretical analysis

The first goal can be summarily filled; tests identical to those described in the numerical integration of this paper were run and produced similar results. The second goal allows us to specify feature selection algorithms that are similar to those used in real fMRI analysis, as well as introduce new feature selection algorithms that have not yet been tested in fMRI experiments. The third goal is the most complicated to define, and will require a choice of the parameters that require exploration. The results from the previous section will help guide the exploration presented in the experiments on synthetic data.

4.2 Generation of Synthetic Data

Synthetic data generation created a dataset similar to an fMRI data set based on supplied parameters. The parameters specified to generate the dataset (from Table 2) were n_f, n_{T1}, n_{T2}, n_{C1}, n_{C2}, n_{C3}, μ_A, μ_{T1}, μ_{T2}, and σ_f. Additionally, the model implemented the assumption that the variance of data is the same regardless of condition or class. Data for each synthetic voxel was drawn independently from a standard normal random generator and scaled to the desired mean and variance.

4.3 Feature Selection in Synthetic Data

The existence of actual data allows greater flexibility for feature selection methods but also requires . For example, while it is possible to select features based on the magnitude of the difference in task means, it is also possible to estimate the variance of the data and use that as a factor in the feature selection process. In addition to the feature selection algorithms that rank voxels based on the difference in task mean and fixation mean for activity tests or the difference between task means for discriminability tests, new algorithms will rank

voxels based on the probability of observing the appropriate difference in means based on the variance of the data. Seeing a high task mean or task difference if there is no expected difference will yield a low probability, so the selection algorithms select the most improbable data given the assumption that there are no discriminative voxels. To supplement the algorithms used in theoretical analysis that simply choose the maximum difference between means, new versions of each algorithm, labeled Activity' and Discriminability', that estimate the variance of the data are introduced below, (note again z is the standard normal PDF $\sim N(0,1)$):

$$\begin{aligned} \Gamma(D, n_V) &= min_V(z(\tfrac{\mu_{T1}-\mu_f}{.5*(\sigma_{T1}+\sigma_f)})) && \text{Activity'} \\ \Gamma(D, n_V) &= min_V(z(\tfrac{|\mu_{T1}-\mu_{T2}|}{.5*(\sqrt{\sigma_{T1}^2+\sigma_{T2}^2})})) && \text{Discriminability'} \end{aligned}$$

In addition to the two algorithms that have been analyzed in detail, tests of synthetic data will include a third feature selection algorithm that will be referred to as three-way feature selection and is slightly more complicated. The idea behind three-way feature selection is to *predict* labels for the training data using parameter estimates from the training data. Specifically, the algorithm (see Table 6) computes voxel-specific means and variances for each condition and then predicts the most likely condition for each data point. Voxels with the greatest accuracy in prediction are judged to be relevant to the task. The major advantages of the three-way feature selection algorithm are that it uses all of the training data in the selection process and it attempts to predict the label of the data, which is similar to the classification task that for which the feature selection algorithm is choosing data.

4.4 Drawbacks of Synthetic Data

The fundamental limitation of synthetic data is the fact that results only pertain to a single sample. In the theoretical analysis, results were obtained from treating the model of the data as a sampling distribution and using the results to create parameter estimates. In the case of synthetic data, samples are generated manually, and running selection on these samples is time consuming. Moreover, the results are derived from a finite number of data sets. As a result, results from synthetic data simulations are noisy, and many samples must be generated to assure performance is consistent. In contrast, synthetic data embodies some characteristics of the real fMRI data in that we have data limited by the number of subjects

```
foreach voxel v
            μ_f = mean( v | cond(v) = fixation);
            σ_f = std( v | cond(v) = fixation);
            μ_1 = mean( v | cond(v) = task 1);
            σ_1 = std( v | cond(v) = task 1);
            μ_2 = mean( v | cond(v) = task 2);
            σ_2 = std( v | cond(v) = task 2);
            foreach time t
                        p_f(t) = z((v(t)−μ_f)/σ_f);
                        p_1(t) = z((v(t)−μ_1)/σ_1);
                        p_2(t) = z((v(t)−μ_2)/σ_2);
                        //(label with the most probable condition)
                        predictedLabel(t) = indexOfMax(p_f(t), p_1(t), p_2(t));
            end
            //(Score voxel based on the accuracy of predicted labels)
            score(v) = labelAccuracy(predictedLabel);
end
```

Table 6: Three Way Selection Algorithm

in an experiment, and the data has noise that may cause the data to deviate from its expected distribution. However, synthetic data also fails to capture all of the details of fMRI data due to the underlying model used to generate the data. As discussed in Section 2:, the model uses a very simple view of fMRI data that assumes that data from voxel populations can be described using three different distributions with equal variances. As a result, synthetic data has some of the strengths of real data as well as theoretical analysis but has properties that prevent it from matching the results of either.

4.5 Results from Synthetic Data Experiments

Experiments from synthetic data will investigate the effect of three parameters on the ability to choose relevant features: the difference in task means, the number of nondiscriminatory signal voxels, and the magnitude of the noise. These three factors may play a very important role in the success of feature selection and the results of these experiments should prove insightful.

The parameters in all of the synthetic data experiments were: $n_{C1} = 3000, n_{C2} = 150, n_{C3} = 50, \mu_A = 2.1, \mu_{T1} = 2.6, \mu_{T2} = 1.5, \sigma_{f,A,T1,T2} = 3, \Gamma = \{$activity', discriminability', three-way$\}$.

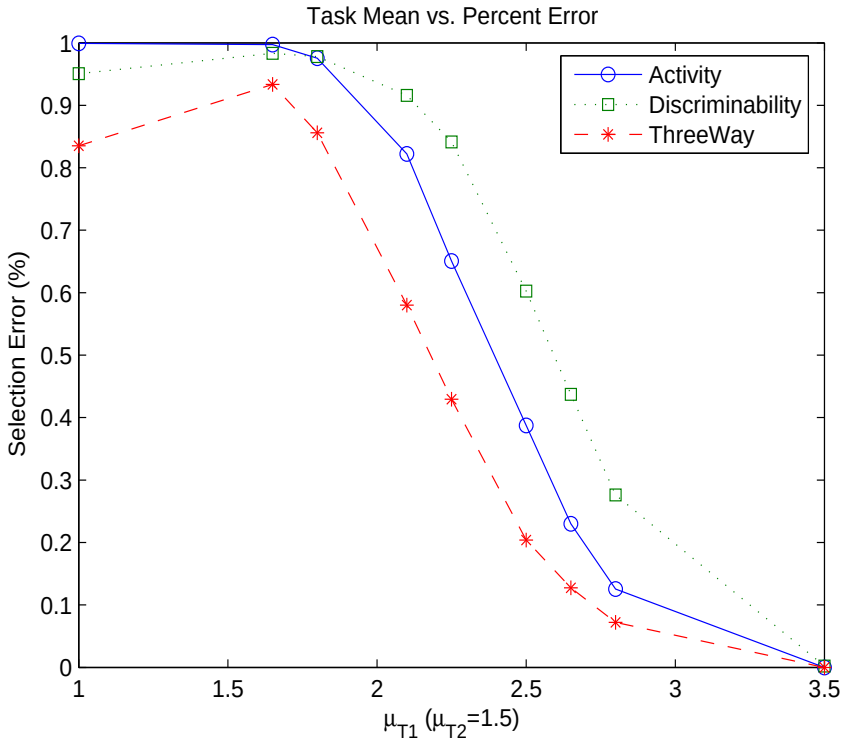

Figure 6: Task 1 Mean and Selection Error (Synthetic Data)

However, in each experiment one of these parameters is varied to view the results. Although the synthetic data experiments used to verify the theoretical analysis used the same parameters as the analysis, different parameters were chosen for synthetic experiments to speed simulation and reduce memory requirements. Results are considered representative of the feature selection problem in general. The results reported in the graphs are percent selection error of feature selection, namely the percentage of selected voxels that are not in Class 3. Figure 6 shows the effect of varying the mean of Task 1, Figure 7 shows the changes that occur as more Class 2 voxels are added, and Figure 8 reveals the effects of noise on feature selection.

The most striking observation that is apparent from the graphs of feature selection is that estimating the variance of the data makes feature selection more difficult, as an additional parameter must be estimated instead of relying on a known value. This effect is particularly pronounced for discriminability tests, which perform more poorly that the other algorithms in almost every condition in each test. For example, discriminability tests are selecting about half of the relevant voxels in Figure 7, even when presented with fewer competitors

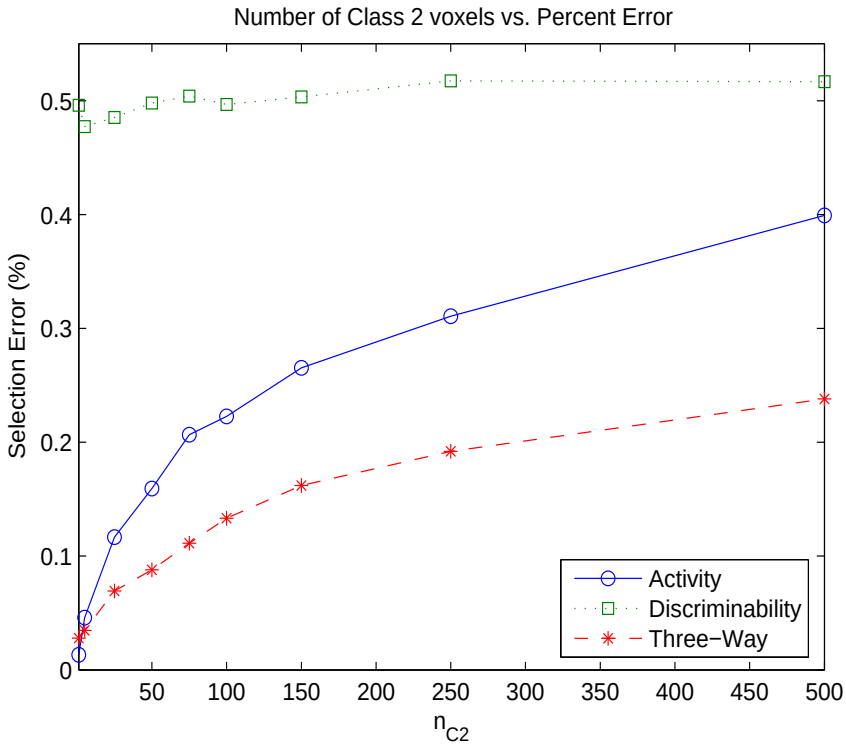

Figure 7: Number of Class 2 voxels and Selection Error (Synthetic Data)

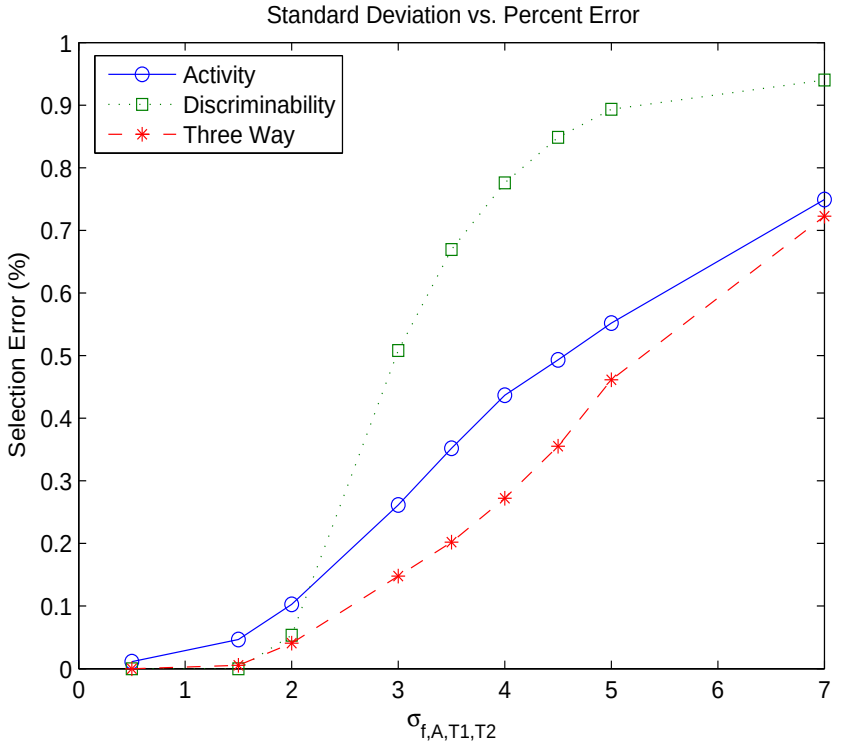

Figure 8: Noise and Selection Error (Synthetic Data)

than in the theoretical analysis, which assumed the presence of 10,000 no-signal voxels. Moreover, it is interesting to see the impact of added noise (Figure 8) on discriminability tests – their error climbs far more rapidly than other tests. Another interesting trend is the sharp drop in error as the difference in task means increases, once again supporting the hypothesis that there is a threshold (see Figure 6, $\mu_{T1} = 2.8$) for the difference in means that allows successful discrimination. Mean differences smaller than the threshold yield very poor selection performance and mean differences greater than the threshold yield good selection performance. The impact of this trend yields a critical insight: if mean differences are too small, discriminability-based feature selection will be unable to discriminate task salience from noise. If the differences in real data are not beyond this threshold, the poor performance of discriminability tests makes more sense.

The contrast in the results of adding nondiscriminatory in these tests is also noteworthy; ten Class 2 voxels were enough to allow discriminability tests to outperform activity tests, but here 150 Class 2 voxels have not caused activity tests to suffer. Perhaps this result speaks to the relative ratios of Class 2 and Class 3 voxels; with ten times as many Class 2 voxels activity test suffer, but with three times as many Class 2 voxels activity tests seem to do fine. This may be another reason that activity tests outperform discriminability tests in this experiment. However, another interesting result is that the new three-way feature selection algorithm outperforms activity tests in every case. As the number of Class 2 voxels increase, the two algorithms appear to approach an asymptote (Figure 7, $n_{C2} = 500$), the difference between the two algorithms appears very significant.

4.6 Conclusions from Synthetic Data

The two questions this study seeks to answer are:

1. Why does activity-based feature selection perform better than discriminability-based feature selection?

2. Can we improve on this performance with another algorithm?

This section has provided answers to both of those questions. Three interesting trends in the experiments with synthetic data support these answers. The first is that the synthetic

experiments show that activity tests outperform discriminability tests in almost every case. Why does this occur? In the theoretical analysis, the feature selection algorithm did not estimate the noise of the data, but in our present analysis the algorithms are more sophisticate and estimate the noise for each voxel. The generated noise has some variance, and the imperfect estimates of noise cause the difference in task means to appear less salient. While a difference of task means of $.4\sigma$ was enough to cause a discriminability test to converge to 0 in the theoretical analysis, the same difference in task means results in feature selection that chooses only half of the discriminative voxels in the data. The observation that a small difference is no longer sufficient for good classification accuracy may be linked to a second observation apparent from the results shown in Figure 8 which demonstrates how sensitive discriminability tests are to noise. Just as we found a separation of means beyond a certain threshold causes the error of discriminability tests to converge to 0, a certain threshold of noise seems to make a large difference to selection accuracy. The final observation is that the Three-Way feature selection algorithm seems to have great potential on the basis of synthetic data experiments; it outperforms activity-based feature selection in all cases studied. This result supports the ability of the new algorithm to perform both a contrast between all conditions, including the fixation conditions. The prediction of fixation data helps the algorithm choose voxels that are truly active, while the contrast between tasks allows it to select discriminative voxels instead of nondiscriminatory active voxels. These results are now evaluated using experimental data.

5: Discoveries from Experimental Data

5.1 Approach

Having discovered some interesting characteristics of feature selection from synthetic data experiments, the next step is to extend the analysis from the previous two sections to real fMRI data. One problem is that the previous sections predicted results that were dependent on the conditions present in fMRI data, however the conditions present in fMRI data are a mystery. Prior to this analysis, it might have been difficult to pinpoint the characteristics of the data that would be useful in determining the nature of fMRI data, but the work thus far suggests some natural questions that an analysis of fMRI data should answer. Specifically,

analysis should seek to answer three questions:

1. What are the characteristics of the overall population of voxels?
2. What are the specific characteristics of individual voxels?
3. How do different feature selection algorithms leverage these characteristics to achieve good performance?

After answering these three questions, the answer to the two big questions that motivated this study, "Why does activity-based feature selection outperform discriminability-based feature selection?" as well as the corollary, "Can we design an algorithm that outperforms activity-based feature selection?" should be apparent. The approach we adopt is to generate histograms of the means, variances and mean differences of the voxels in the brain in aggregate, and then take a more detailed look at the voxels selected during activity, discriminability, and three-way tests in hopes of gaining a better understanding of the operation of each of these algorithms. Before proceeding with this discussion, a few prefacing remarks about the data set are necessary.

5.2 Experimental Description

The data set used in this analysis investigates a semantic categories in five subjects, and was provided by Professor Marcel Just at Carnegie Mellon University. In this experiment, subjects are asked to classify each word in a series based on whether it belongs to a specific category. The three categories in this study were fish, vegetables, and trees. The experiment was organized into a series of presentation sets, each of which consisted of trials that contained blocks of words. There were a total of four presentation sets, each of which consisted of three trials, for a total of twelve trials. The presentations of categories were counterbalanced to avoid deterministic interference effects. Each of the trials contained a block of 20 words from one of the categories. Each block of words consists of twenty repetitions of a word presentation lasting 400 milliseconds followed by 1200 milliseconds of a fixation screen with an 'x' at the center. Since there were twenty repetitions of a 1.6 second task, the data per trial consists of 32 images. Data was collected for the fixation condition from 24-second fixation periods that were present at the beginning and end of the experiment as well as after the fourth and eighth trials.

Beyond pre-processing at the collection center, the data was temporally convolved with a smoothing kernel ([.1 .2 .4 .2 .1]) on a per-voxel basis, and each presentation set was normalized by subtracting the mean activation over the entire presentation set from each image. Separate training and testing sets were used for the feature selection and classification process, although the details differed based on the type of experiment. The analysis included "primitive experiments" that were computationally inexpensive and "elaborate" experiments that increased the amount of cross-validation but required more computation. Primitive experiments used the third presentation set as the test set, used the entire brain for feature selection, and classified each image from the test set using Gaussian Naive Bayes or K-Nearest Neighbor classification. More involved experiments restricted data to selected regions of interest, specifically the occipital pole (OP), calcine fissure (CALC), bilateral inferior extrastriate cortex (LIES,RIES), bilateral inferior temporal cortex (RIT,LIT), and bilateral temporal cortex(RT,LT). These experiments performed Leave-3-Out cross validation for feature selection and classification, and created an average image for each trial for the classification step. Since there are four presentations of each category, there are $4^3 = 64$ different ways to exclude a trial from each of the three categories, as a result the accuracies reported are the average number of correct classifications from these 64 different combinations of data used to form training sets. Since trials had been converted to average images, fewer training images were available and the K-Nearest Neighbor classification algorithm seemed most appropriate for analysis.

5.3 Additional Feature Selection Algorithms

Some extensions to the three algorithms discussed so far will also be tested during experimental trials. The first two algorithms are layered feature selection algorithms based on the core activity and discriminability methods that involve using one feature selection algorithm (i.e. an activity-based test) to select s voxels, and then from the pool of s voxels choose the desired n_V voxels using the other feature selection algorithm (i.e. discriminability). These two algorithms will be referred to as Layered-AD and Layered-DA, where the last two letters specify the order of the feature selection algorithms. Two extensions to the three-way selection algorithm that have been widely adopted in other scenarios by machine learning researchers also merit discussion. The first extension is *spatial pooling* where the selection

algorithm assumes all voxels have the same variances, although these variances are computed separately for each experimental condition. The second extension is *temporal pooling* where the selection algorithm assumes the variance of a voxel is the same for the entire time course regardless of the experimental condition, although different voxels still have different variances. By using these simplified variance estimates, additional parameters to the model can be removed reducing the potential for overfitting.

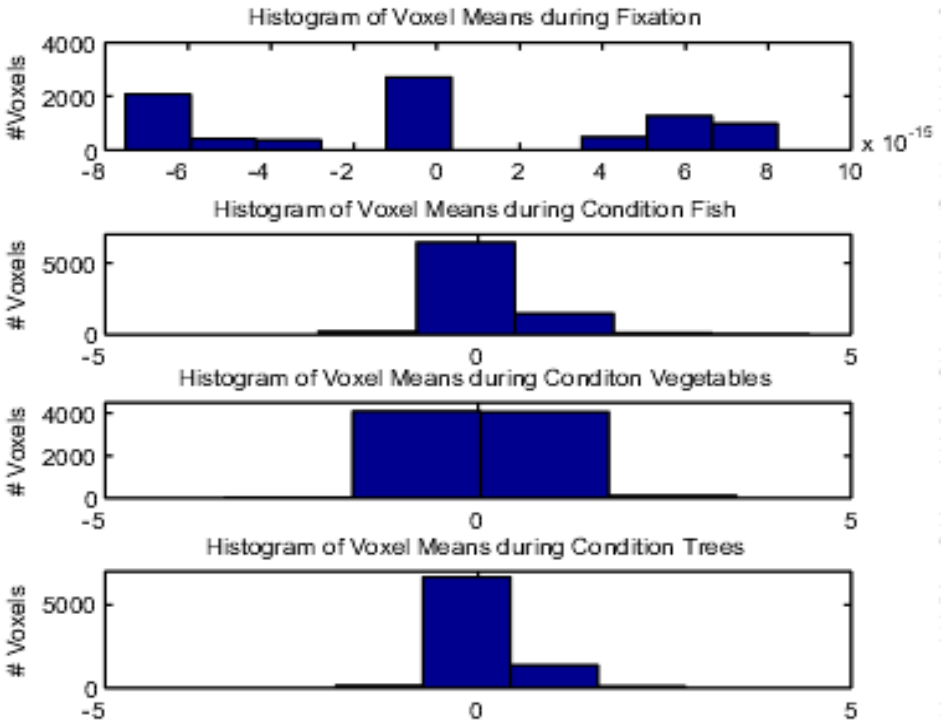

Figure 9: Histogram of Voxel Mean Values

5.4 Profiling fMRI Data

As mentioned earlier, fMRI data was profiled using histograms of means, variances, and mean differences over the entire population of voxels. Figure 9 shows the distribution of means for the four possible conditions (fixation, fish, vegetables, trees). Notice that the x-axis for the fixation condition is in the scale of 10^{-15}. It seems reasonable to assume that no voxels were very active during fixation. Moreover, the bulk of voxels have low task means in the range of (-2,2), while a select few have higher means. Figure 10 shows only larger means. It might be reasonable to assume that these voxels are the ones that are most

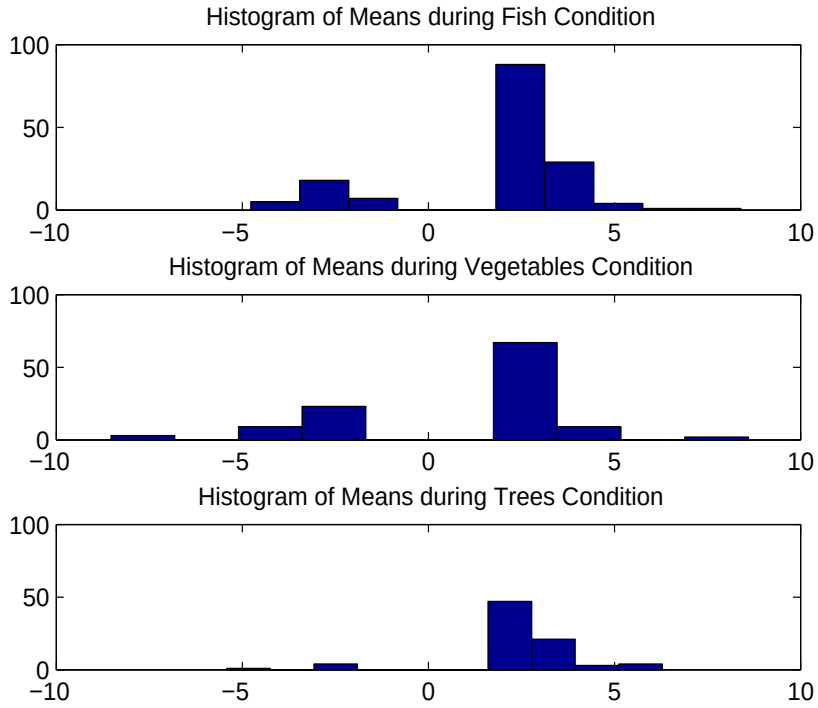

Figure 10: Histogram of large Mean Values

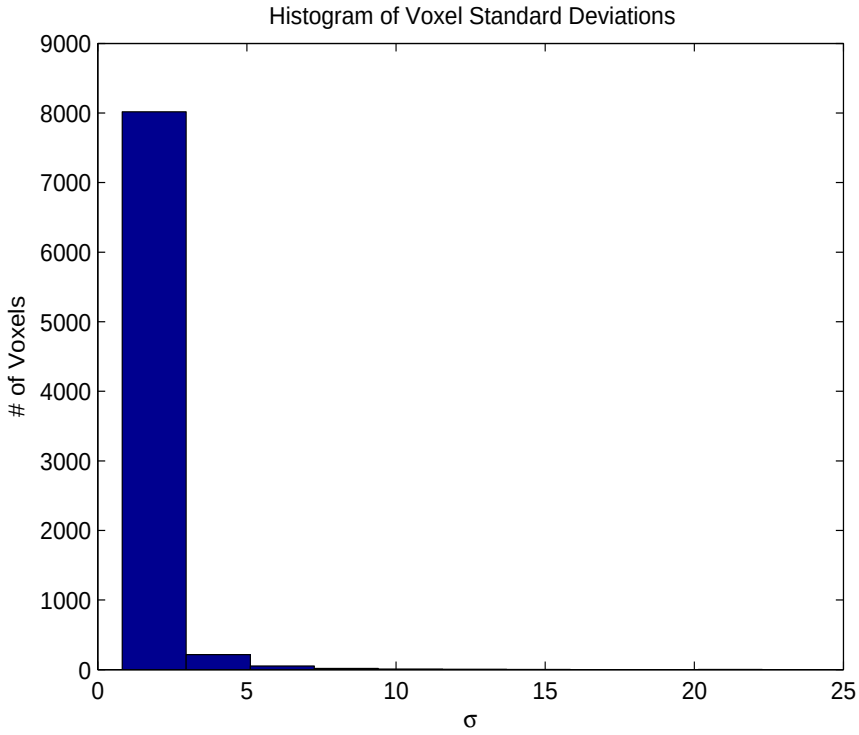

Figure 11: Histogram of Voxel Standard Deviations

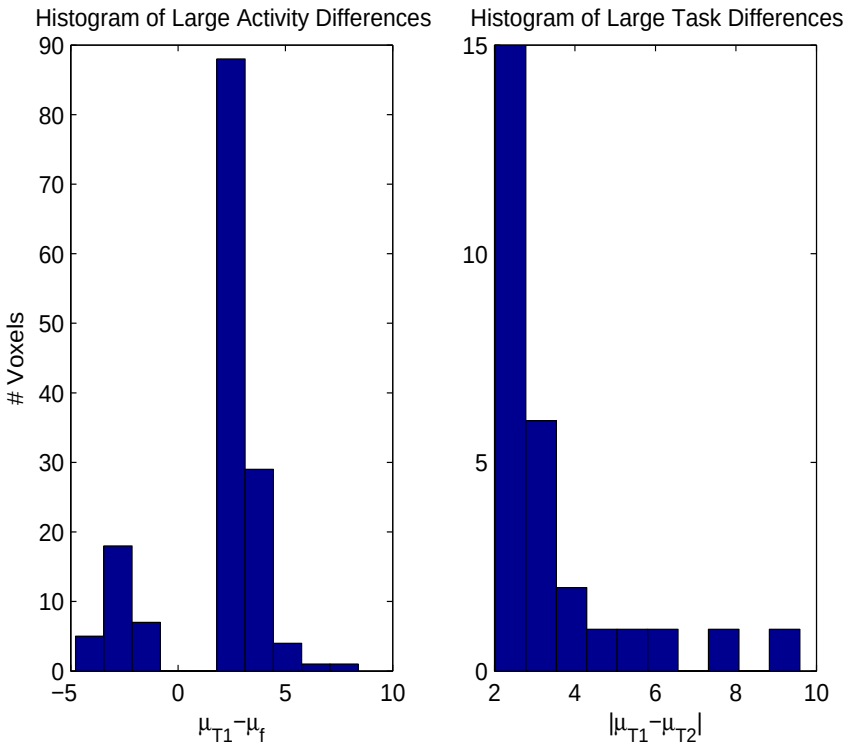

Figure 12: Histogram of Large Activity/Task Differences

discriminative. Figure 11 shows the distributions of the standard deviation over the entire time course for the population of voxels. Note that most standard deviations are between 0 and 2.5, with a small fraction larger than 2.5. Since task means are small, it made sense to show a histogram of large task-activity differences and task-task differences. Figure 12 compares the large differences in means between fish and fixation as well as between fish and vegetables, removing differences in the range of (-2,2). The scale of the histogram confirms that only a few voxels are either active or discriminatory for a given contrast. It is also readily apparent that more voxels are active, while fewer voxels have large differences between means. Is this grounds to conclude that activity-based feature selection may do better than discriminability-based feature selection simply on the basis of having more voxels to choose? The hypothesis is certainly interesting, and will be investigated in the next section.

5.5 Profiling Feature Selection

Having compiled aggregate statistics for the population of the brain, it seemed that the next step would be to take a look at "case studies" in feature selection - voxels that are ranked

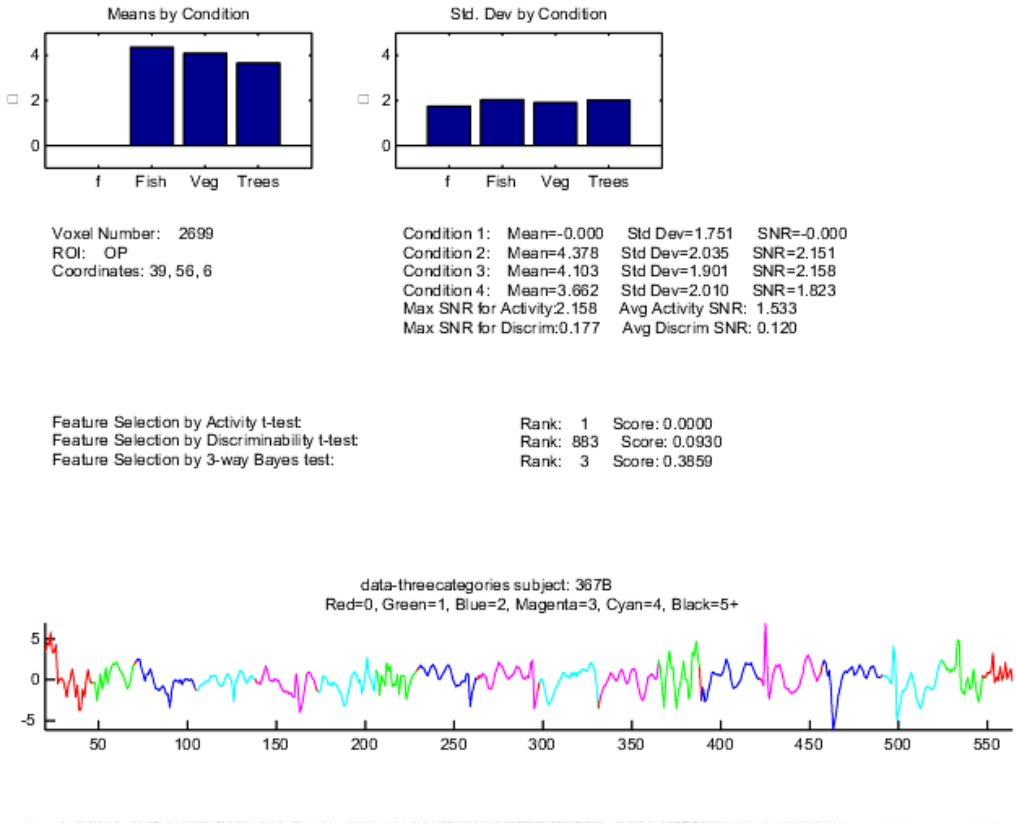

Figure 13: Highest Scoring Voxel in an Activity Test

as very discriminative and somewhat discriminative by the feature selection algorithms. To achieve this end, we created a function that would take the scores of voxels from the three different feature selection algorithms and create "profiles" of the voxels predicted to be most discriminative, including information such as the feature score, anatomical region, mean, variance, and time course of the data in the profile.

Consider the profile for the top voxel selected by activity-based feature selection, as shown in Figure 13. Note that this voxel is also scored highly by a three-way test, although not by a discriminability test. Although the differences in the means are not very large, the means themselves are high and the standard deviation is also fairly low, yielding a high signal-to-noise ratio. Note that the standard deviation is approximately equal across conditions. Contrast this profile with that of the best voxel according to a discriminability test.

Figure 14 shows the top voxel as ranked by discriminability scoring. Notice that the other two algorithms do not rank this voxel highly, although the three-way selection algorithm

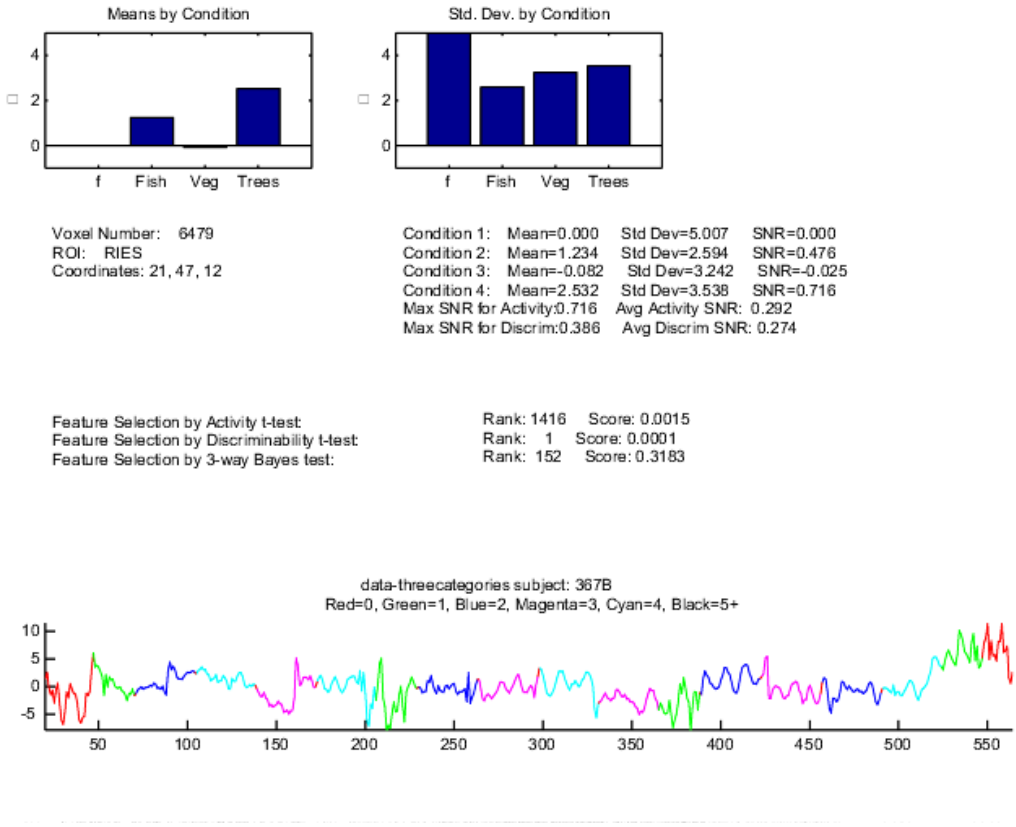

Figure 14: Highest Scoring Voxel in a Discriminability Test

assigns a much better ranking to this voxel than the activity-based algorithm. Moreover, note that the differences in mean value are fairly large, close to the estimates used in the modeling simulations in previous sections as well as large relative to one another. However, the standard deviations vary greatly depending on the experimental condition and are all fairly large, especially in comparison with the standard deviation of the voxel selected by activity-based selection. Paired with the disastrous effect of a high standard deviation shown in the synthetic data studies, this pairing suggests the fundamental problem with discriminability based feature selection is a low signal-to-noise ratio. How can the desire for discriminative voxels be reconciled with robust feature selection? Three-way feature selection may be the answer.

Figure 15 shows the best voxel in the three-way selection ranking system. Note that this voxel is ranked highly by both the activity-based *and* the discriminability based selection methods. Moreover, the difference in the mean values is relatively large, although

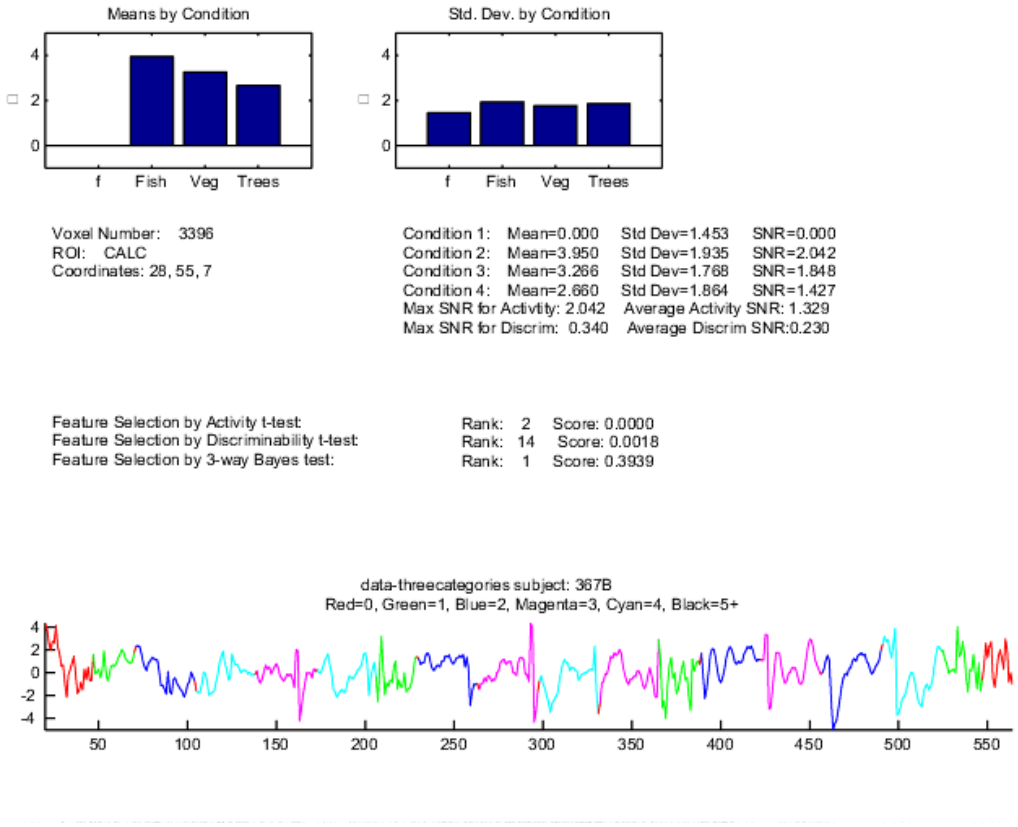

Figure 15: Highest Scoring Voxel in a Three-Way Test

not quite as large as the values used in our modeling. Compensating for the smaller differences in mean are lower standard deviations, which are close to those selected by the top voxel activity-based feature selection. The standard deviations are also relatively consistent between conditions, unlike the variances in the voxel of the discriminability test. For the first 25 features, three-way selection consistently chooses high signal-to-noise voxels, and the choice of these voxels seems to influence its success in synthetic experiments.

Given the overlap between the three-way selection algorithm and the two other algorithms, it would be interesting to quantify the overlap. Instead of simply reporting an overlap, a more insightful idea might be investigating the anatomical regions that overlap between feature selection algorithms. Figure 16 demonstrates the overlap between three-way selection and activity and discriminability tests. The three-way algorithm selects voxels from the same regions as both activity and discriminability tests, while some regions show no overlap whatsoever between activity and discriminability tests. Note that a total of 17 voxels are

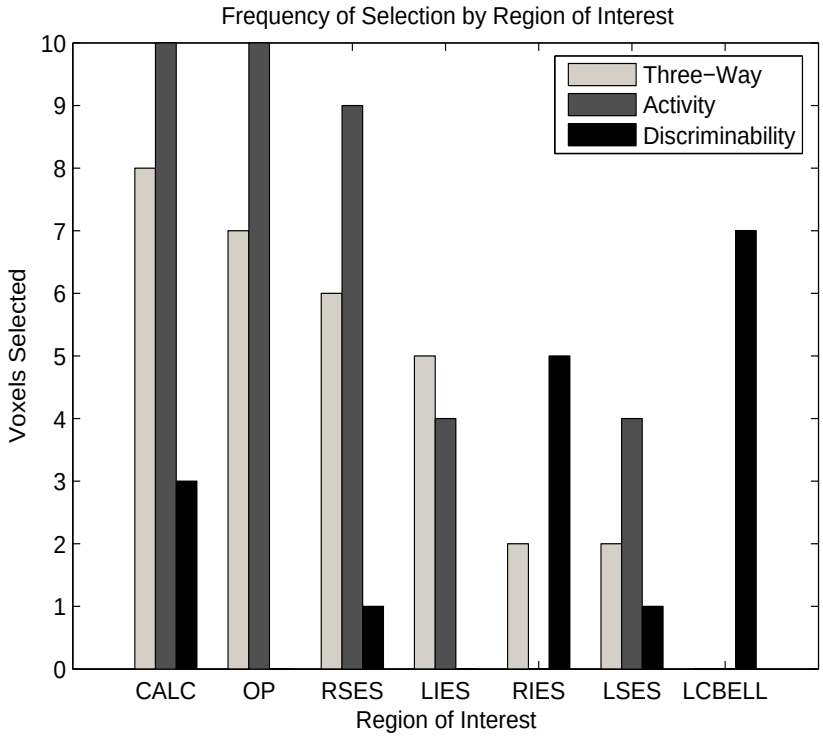

Figure 16: Frequent Regions of Feature Selection by Algorithm

chosen from the region "OP" (occiptal pole) between activity and three-way selection tests while none are selected by discriminability. Conversely, both discriminability and three-way selection tests select data from "RIES" (right inferior extrastriate cortex), while activity-based selects no voxels from the region. Finally, discriminability tests can choose data from regions that may be very discriminative but not necessarily relevant to the task. For example, discriminability-based selection chooses seven voxels from the region "LCBELL" (left cerebellum) while no other algorithms choose from that region. Also note that the regions displayed are those with the most voxels represented by selection. Looking at the entire frequency list shows that voxels selected using activity-based methods are clustered in a few regions, while discriminability-based selection chooses a few voxels from many regions in the brain. Three-way selection compromises between the two, still selecting clusters of voxels from some regions but selecting from more regions overall than activity-based selection.

While three-way selection seems to incorporate the best of both worlds, the profiles show that the performance of the algorithm may begin to degrade quickly. Activity-based tests continue to pick voxels with low variance with respect to task means; the standard deviation

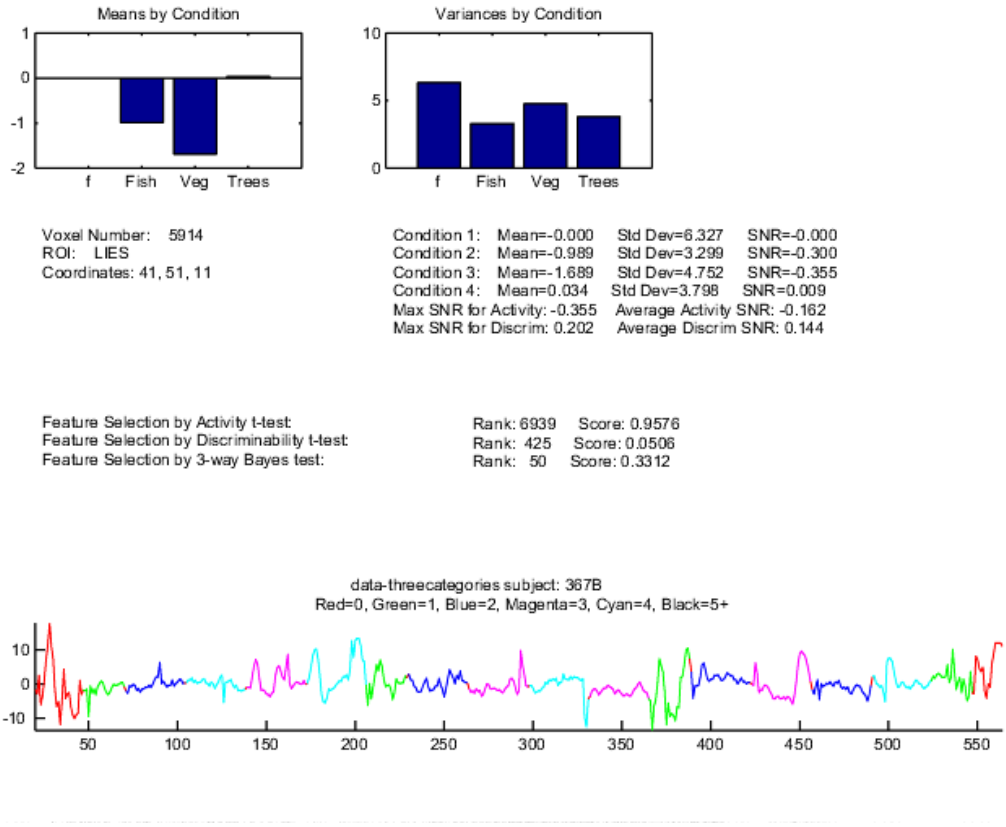

Figure 17: Low Signal-to-Noise Ratios in Three Way Feature Selection

is less than or approximately equal to the task means in the top fifty selected voxels. Discriminability tests continue to choose discriminative voxels even if they have high variance, in the worst case the standard deviation may be three times the difference in task means. Contrast this with three-way selection, which selects voxels that show a trend toward increasing variance with respect to mean, yielding lower signal-to-noise ratios as more voxels are chosen. While the standard deviations remain low, remaining in the range of 1 to 3, the task means can sometimes me small as well. Figure 17 shows the state of three-way feature selection at the end of the selection set, with very low signal-to-noise ratios. The problem in this scenario is that, as noted earlier, discriminability measures can be excessive as the difference in means need not be very large to obtain good classifier accuracy. On the other hand, the effects of variances cause performance to degrade much more rapidly. As a result, it seems likely that three-way selection will suffer the same fate as discriminability-based feature selection for larger selection sets, although cursory investigation shows no obvious trend in the deterioration of classification accuracy.

In contrast, the layered feature selection algorithms described earlier intuitively show some promise. If activity-based selection is capable of selecting consistent, low-variance voxels then it should be able to select the best voxels from a seed pool of discriminative voxels. Correspondingly, if activity-based feature selection can guarantee low-variance voxels, then a discriminability test will find the most discriminative voxels in that pool. The problem is that even in the best voxels selected by these two approaches, the ranking for the complementary algorithm is fairly low. It's possible that the voxels an activity-test would choose might never show up in the top 10% of voxels selected using a discriminability-test.

5.6 Results from Classification

Γ / Subject	354B	357B	362B	367B	371B	Average
Activity'	.245	.294	.598	.461	.560	.431
Discriminability'	.392	.255	.461	.324	.464	.380
Layered-AD	.402	.226	.412	.392	.310	.348
Layered-DA	.304	.245	.441	.461	.610	.412
ThreeWay	.382	.284	.500	.510	.369	.410
ThreeWay-Spatial	.431	.147	.589	.216	.464	.365
ThreeWay-Temporal	.402	.265	.461	.490	.417	.407

Table 7: Naive Bayes Classification Accuracy for Five Subjects

The culmination of this thesis is to attempt to apply the findings from all of the earlier sections in the form of implementation and experimental validation. Tests of synthetic data showed that three-way feature selection would outperform activity-based feature selection. By applying a three-way feature selection algorithm to the three-categories data set to choose

Γ / Subject	354B	357B	362B	367B	371B	Average
Activity'	.255	.422	.324	.314	.490	.361
Discriminability'	.490	.382	.294	.196	.422	.357
Layered-AD	.324	.343	.333	.451	.255	.341
Layered-DA	.304	.284	.431	.441	.411	.376
ThreeWay	.412	.226	.402	.304	.382	.345
ThreeWay-Spatial	.422	.275	.373	.422	.402	.337
ThreeWay-Temporal	.245	.411	.441	.235	.252	.378

Table 8: 1-Nearest Neighbor Accuracy for Five Subjects

fifty voxels and attempting to classify the test set on the basis of those features, the resulting accuracies may support the use of three-way feature selection.

Note that the first set of data is from "primitive experiments", where presentation set 3 was reserved as a test set and the remaining data was used for feature selection and classifier training. An algorithm that randomly guessed labels would have an accuracy of .333. The accuracy of classification is shown in Table 7 for the Gaussian Naive Bayes classifier and Table 8 for a 1-Nearest Neighbor classifier. The feature selection algorithms tested are those mentioned in the "Feature Selection Algorithms" section of the synthetic data. Briefly reviewing these algorithms:

Activity': Assumes that there are no active voxels and ranks voxel data based on how "improbable" it is based on this assumption.

Discriminability': Assumes that there are no discriminative voxels and ranks voxel data based on how "improbable" it is based on this assumption.

Layered-AD: Uses the Activity' test to select 10% of the data as a selection pool and then chooses the requisite voxels from the selection pool using Discriminability'

Layered-DA: Uses the Discriminability' test to select 10% of the data as a selection pool and then chooses the requisite voxels from the selection pool using Activeity'

ThreeWay: Described in Table 6. Estimates the parameters of the distributions and ranks voxels based on how well their parameters predict the trial labels.

ThreeWay-Spatial: Similar to ThreeWay, but uses the variance estimate for the entire brain for each condition, so in the case of three conditions there are three estimates for the entire brain.

ThreeWay-Temporal: Similar to ThreeWay but assumes the variance of a voxel is the same for all conditions.

From the results of the Naive Bayes classification it appears that no variant of ThreeWay or Layered clearly outperforms activity, and the results of 1-Nearest Neighbor classification show a similar trend, although using the Layered-DA algorithm does increase the accuracy of the Activity test. The failure of these algorithms throws suspicion on the model of data used in the previous sections; if the predictions regarding the relative performance of feature

Γ / Subject	354B	357B	362B	367B	371B	Average
Activity' + 1-NN	.406	.552	.583	.974	.802	.663
ThreeWay + 1-NN	.833	.521	.667	.917	.833	.754
Activity' + 2-NN	.375	.375	.604	.990	.698	.608
ThreeWay + 2-NN	.719	.349	.604	.917	.771	.672
Activity' + 3-NN	.370	.443	.656	.958	.750	.635
ThreeWay + 3-NN	.833	.344	.620	.917	.833	.695

Table 9: K-Nearest Neighbor Accuracies with Leave-3-Out Cross-Validation

selection algorithms were incorrect, perhaps the other insights gained from the model are also subject to inaccuracy in the face of variable data. Given these disappointing results, a more elaborate training-testing cross-validation procedure mentioned earlier, that used k-Nearest Neighbor on averaged images was used to confirm this conclusion. The results from a Leave-3-Out cross-validation are shown in Table 9. Notice that rigorous testing seems to show a clearer advantage for the three-way feature selection algorithm regardless of the number of neighbors. Moreover, the accuracy of this algorithm across subjects is fairly impressive, achieving over 75% average accuracy in comparison with a 33% baseline. Although the reconciliation between these results and those presented from the more primitive testing method is not obvious, the proposed three-way feature selection algorithm clearly shows potential in the problem setting described here, and merits further study.

6: Conclusion

This document describes a project devoted to understanding an apparent contradiction in feature selection, the superior performance of activity tests in classification tasks. Creating a model of fMRI data and defining the feature selection process analytically permitted a glimpse into the factors that influence feature selection. One puzzle that arose was the interference presented by nondiscriminatory active voxels in activity-based feature selection. Some hypotheses to account for this discrepancy were tested, but the most fruitful results came from experiments with synthetic data.

Initial experiments using synthetic data confirmed the findings presented by the analytical model, namely that discriminability tests outperform activity tests in most cases due to the ability of discriminability tests to perform accurate feature selection with a small difference

in task means despite significant noise. More elaborate experiments using different feature selection methods demonstrated that the magnitude of the difference in task means coupled with high variance could have severely adverse effects on discriminability-based feature selection. In addition to this finding, synthetic data suggested a new feature selection algorithm that predicted the labels, or experimental condition, of all data could outperform both discriminability and activity-based feature selection. This hypothesis was investigated and tested using real fMRI data.

The approach to understanding real fMRI data involved profiling the activity and variability of the entire population of voxels in the brain, and then concentrating on the details of the top voxels chosen by our feature selection algorithms. The findings from aggregate studies showed that most voxels in the brain are either inactive or not very discriminative. Narrowing the search to voxels showing activity or discriminability showed a small subset of the brain that demonstrated high activity or high discriminability. In hopes of finding more information about these voxels, profiles of the top voxels from feature selection were scrutinized for clues about the underlying operation of feature selection.

The data seem to suggest that different feature selection algorithms select based on different metrics. Activity-based feature selection selects voxels that show task activity but since many voxels fit this description, those with the lowest variance are chosen. Discriminability-based feature selection chooses voxels that discriminate well between tasks, but will select voxels with high variance if they are discriminative enough. The new candidate feature selection algorithm, three-way feature selection, attempts to straddle the gap between the two major feature selection algorithms. The voxels it chooses are both discriminative and active. The initial choices from this algorithm have low variance, but as more selections are made, low variance is sometimes compromised for high discriminability.

If there is a key to understanding feature selection in fMRI data, the most likely keyword is variance. The reason discriminability tests perform poorly in a variety of situations is that they select discriminative voxels at the cost of data variance. High variance implies that the voxel may have inconsistent activity, but can also cripple a feature selection algorithm that uses estimates of variance such as the Gaussian Naive Bayes classifier. On the other hand,

the performance of activity-based feature selection takes no large risks. Voxels selected may barely discriminate between conditions, but the data itself has low variance with respect to the means. This, it seems, is a recipe for success; if the chosen voxel is not very discriminative, its contribution in classification is unnoticed while the occasional discriminative voxel forms the heart of the classification algorithm.

To supplement these investigations, experimental tests of the different feature selection algorithms were undertaken on a semantic categories data set. While the results do not entail a significant difference between feature selection algorithms, data from other studies or more subjects might clarify the conclusions from this study. Additionally, extensions to the feature selection algorithms described here may have the potential to produce selection algorithms that have better performance for the classification of the "cognitive state" of subjects as well as greater anatomical or functional relevance.

Bibliography

[Cox and Savoy, 2003] David D. Cox and Robert L. Savoy. Functional magnetic resonance imaging (fmri) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *NeuroImage*, 19(2):261–270, June 2003.

[Eagle, 2002] Nathan Eagle. Feature selection analysis. Technical Report TR-557, MIT Media Lab Vision and Modeling, 2002.

[Friston *et al.*, 1995] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C. Frith, and R.S.J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[Mitchell *et al.*, 2003] T. Mitchell, R. Hutchinson, M. Just, R. Niculescu, F. Pereira, and X. Wang. Classifying instantaneous cognitive states from fmri data, 2003.

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