Analysis of Functional MRI Data Using Mutual Information *

Andy Tsai[1], John W. Fisher III[1,2], Cindy Wible[3,4], William M. Wells III[2,3], Junmo Kim[1], and Alan S. Willsky[1]

[1] Laboratory for Information and Decision Systems,
Massachusetts Institute of Technology,
Cambridge, MA, USA
{atsai, junmo, willsky}@mit.edu
http://ssg.mit.edu/
[2] Artificial Intelligence Laboratory,
Massachusetts Institute of Technology,
Cambridge, MA, USA
{fisher, sw}@ai.mit.edu
http://ai.mit.edu/
[3] Department of Radiology,
Brigham and Women's Hospital,
Harvard Medical School,
Boston, MA, USA
cindy@bwh.harvard.edu
http://splweb.bwh.harvard.edu:8000/
[4] Department of Psychiatry,
Brockton/West Roxbury VAMC,
Harvard Medical School,
Brockton, MA, USA

Abstract. A new information-theoretic approach is presented for analyzing *f*MRI data to calculate the brain activation map. The method is based on a formulation of the mutual information between two waveforms– the *f*MRI temporal response of a voxel and the experimental protocol timeline. Scores based on mutual information are generated for all voxels and then used to compute the activation map of an experiment. Mutual information for *f*MRI analysis is employed because it has been shown to be robust in quantifying the relationship between any two waveforms. More importantly, our technique takes a principled approach toward calculating the brain activation map by making few assumptions about the relationship between the protocol timeline and the temporal response of a voxel. This is important especially in *f*MRI experiments where little is known about the relationship between these two waveforms. Experiments are presented to demonstrate this approach of computing the brain activation map. Comparisons to other more traditional analysis techniques are made and the results are presented.

* This work was supported by ONR grant N00014-91-J-1004 and by subcontract GC123919NGD from Boston University under the AFOSR Multidisciplinary Research Program on Reduced Signature Target Recognition.

1 Introduction

We present a novel method based on an information-theoretic approach to find the brain activation maps for *f*MRI experiments. In this method, mutual information is calculated between the temporal response of a voxel and the protocol timeline of the experiment. This value can then be used as a score to quantify the relationship between the two waveforms. Mutual information is appropriate for *f*MRI analysis because it has been shown to be more robust than other methods in identifying complex relationships (i.e. those which are nonlinear and/or stochastic). More importantly, our nonparametric estimator of mutual information requires little *a priori* knowledge of the relationship between the temporal response of a voxel and the protocol timeline. Over the past few years, mutual information has been used to solve a variety of problems [2, 8, 9].

2 Background

2.1 Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging is a powerful new imaging modality with the ability to noninvasively generate images of the brain that reflect brain tissue hemodynamics. Brain tissue hemodynamics are spatially related to the metabolic demands of the brain tissue caused by neuronal activity. Therefore, indirectly, this imaging modality can capture brain neuronal dynamics at different sites while being activated by sensory input, motor performance, or cognitive activity.

The specific area of *f*MRI analysis we address in this paper is the identification of those voxels in the *f*MRI scan which are functionally related to the experimental stimuli. This entails determining whether the acquired temporal response of a voxel during the scan is related to the experimental protocol timeline that is used during the scan. This relationship is difficult to establish for the following reason: it is known from single unit recording studies that the response characteristics of neurons differ between brain regions *and* in relationship to different stimuli. Some neurons may respond to stimuli with brief transient activity, whereas others might show more sustained activity to the same stimulation. As cognitive and psychological variables such as habituation and attention are added to the equation, the relationship between brain activity and stimuli becomes even more complex [7]. This, coupled with the fact that *f*MRI measurements–which do not directly measure brain activities–are many steps removed from single unit recordings, makes the relationship between the two waveforms even harder to establish. Because of the complex, most certainly nonlinear and perhaps stochastic, nature of the relationship between the two waveforms, it has been difficult to find a suitable metric to quantify the dependencies. The technique we present in this paper can be used to overcome such obstacles.

2.2 Popular Strategies for Analysis of *f*MRI data

Currently, the popular analysis methods used to obtain the activation map includes direct subtraction [5], correlation coefficient [1, 10], and the general linear

model [4]. Quantitative comparisons of these methods are difficult given the absence of ground truth, little knowledge about human brain activation patterns, and the indirect role *f*MRI plays in capturing brain activation. The following is a short description of the popular *f*MRI analysis techniques.

Direct Subtraction (DS) This method involves calculating two mean intensities for each voxel–one mean value calculated based on averaging together all the temporal responses acquired during the "task" period, and the other mean value calculated based on averaging together all the temporal responses acquired during the "rest" period of an experiment. To determine whether a voxel is activated or not, one mean intensity is subtracted from the other. Voxels with significant difference in the mean intensities of the two data groups are identified as being activated. To yield a statistic to identify significant difference in the intensities, a Student's t-test is employed. This test determines whether the means of the two data groups are statistically different from one another by utilizing the difference between the means relative to the variabilities of the two data groups. The t-value this method generates, for a temporal response y, is calculated as

$$t = \frac{\bar{y}_{on} - \bar{y}_{off}}{\sqrt{\frac{\sigma^2_{y_{on}}}{N_{on}-1} + \frac{\sigma^2_{y_{off}}}{N_{off}-1}}}$$

where y_{on} and y_{off} denote the set of data points in the temporal measurements that correspond to the "task" and the "rest" periods, respectively, and N_{on} and N_{off} denote the number of time points that corresponds to the "task" and the "rest" periods, respectively. The mean and variance of the data group y_{on} are denoted as $\bar{y}_{on}$ and $\sigma^2_{y_{on}}$, respectively. Likewise, the mean and variance of the data group y_{off} are denoted as $\bar{y}_{off}$ and $\sigma^2_{y_{off}}$, respectively. The major shortcoming associated with this method is that it relies heavily on the assumption that temporal measurements of a given voxel can be partitioned into two data groups, each normally distributed according to a different mean and variance.

Correlation Coefficient (CC) The correlation coefficient ρ_{xy} is a normalized measure of the correlation between the reference waveform x and the measurement waveform y, and is defined by

$$\rho_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ denote the means of x and y, respectively. The summation is taken over all the time points in the waveform. It is easy to establish that $-1 \leq \rho_{xy} \leq 1$. Voxels with large $|\rho_{xy}|$s are considered to be activated. For this method, $|\rho_{xy}|$ is used as the test statistic for statistical inference. Given the design of this method, it is expected that the choice of the reference waveform is vital to the success of this technique. Various waveforms have been used [1,10], however, with so many unknown factors at play in measuring the brain activation patterns, it is difficult to pin point which reference waveform is the appropriate one to use.

General Linear Model (GLM) The statistical models used for parameter modeling in the two previously described analysis methods are both special cases of the general linear model. This model is a framework designed to find the correct linear combination of explanatory variables (such as hemodynamic response, respiratory and cardiac dynamics) that can account for the temporal response observed at each voxel during an experiment. Assume that there exists T number of time point measurements per voxel in the fMRI data set. Let y_t denote the measurement at some voxel at time t, and let ϵ_t denote the error term associated with the linear model fit at that same voxel at time t, with $1 \leq t \leq T$. Here, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Suppose there are J number of explanatory variables in the linear model. Let x_{jt} denote the value of the jth explanatory variable at time t with $1 \leq j \leq J$. Also let β_j denote the scaling parameter for the jth explanatory variable. With these definitions, the general linear model can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1J} \\ x_{21} & x_{22} & \dots & x_{2J} \\ \vdots & & \ddots & \vdots \\ x_{T1} & x_{T2} & \dots & x_{TJ} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_J \end{bmatrix}.$$

The above equation can be written succinctly in matrix notation as $Y = X\beta + \epsilon$. In general, X is full rank and the number of explanatory variables J is less than the number of observations T indicating that the method of least squares can be employed to find the scaling parameters β. Since X^TX is invertible, the least squares estimate for β, which we denote by $\hat{\beta}$, is $(X^TX)^{-1}X^TY$. Then $\hat{\beta}$ is used to test whether it corresponds to the model of an activation response (as specified in X) or the null hypothesis. One of the major problems associated with this method is in the design of X. As mentioned earlier, little is known about the relationship between fMRI temporal response and brain stimulation. Hence, it is difficult to identify the necessary explanatory variables that can account for the temporal responses seen in fMRI measurements.

3 Description of Method

3.1 Mutual Information and Entropy

Mutual information (MI) and entropy are concepts which underly much of information theory [3]. They cannot be adequately described within the scope of this paper. Suffice it to say that MI is a measure of the information that one random variable (RV) conveys about another, and entropy is a measure of the average uncertainty in a RV. Both quantities are expressed in terms of bits of information. Here, we demonstrate the appropriateness of MI for fMRI analysis.

The mutual information, $I(u, v)$, between the RVs u and v, is defined as [3]

$$I(u, v) = h(v) - h(v|u) = h(u) - h(u|v), \tag{1}$$

where the entropy, $h(v)$, quantifies the randomness of v and the conditional entropy, $h(v|u)$, quantifies the randomness of v conditioned on observations of

u. These terms are described by the following expectations:

$$h(v) = -E_v\left[\log_2(P(v))\right]$$
$$h(v|u) = -E_u\left[E_v\left[\log_2(P(v|u))\right]\right].$$

where P denotes probability density. It is clear from (1) that MI is symmetric. That is, the information that u conveys about v is equal to the information that v conveys about u. Furthermore, since u is a discrete RV in our case and conditioning always reduces uncertainty ($h(u|v) \leq h(u)$), v can convey at most $h(u)$ bits of information about u (and vice versa). We can therefore lower and upper bound the MI between u and v by 0 and $h(u)$, respectively.

3.2 Calculation of Brain Activation Map by MI

We present nonparametric MI as a formalism for uncovering dependencies in calculating the *f*MRI activation map. Recall that in our specific application, we seek at most one *bit* of information (whether or not a voxel is activated). This impacts our choice of the reference waveform. The reference waveform need be no more complicated than our hypothesis space (1 bit). The protocol timeline shown in Fig. 1 is the simplest model of our hypothesis space and is sufficient as the reference waveform when using MI as the basis for comparison. More elaborate waveforms can be employed, but they imply more information than is necessary. The consequence of this is that complicated waveform design in unnecessary; the reference waveform need only adequately *encode* the hypothesis space.

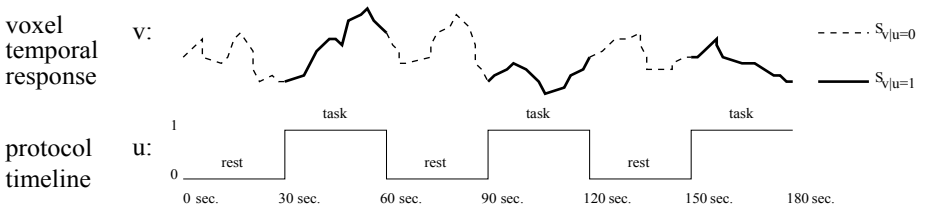


Fig. 1. Illustration of the Protocol Timeline, $S_{v|u=0}$, and $S_{v|u=1}$.

In the following derivation, we will refer to the temporal response of a voxel as v, and the reference waveform as u. We have already established the appropriateness of using the protocol timeline as the reference waveform u. As such, u only takes on two possible values, 0 and 1, so we can rewrite equation (1) as

$$I(u,v) = h(v) - P(u=0)h(v|u=0) - P(u=1)h(v|u=1) \qquad (2)$$

where $P(u=0)$ and $P(u=1)$ are the *a priori* probabilities of u taking on the values of 0 and 1, respectively. By the nature of the protocal timeline, these two probabilities are both 0.5.

As an illustrative example, suppose v is a scaled and biased version of u (i.e. $v = cu + d$ where $c, d \in \Re$ and $c \neq 0$). Then

$$h(v|u=0) = -E_v\left[\log_2(P(v|u=0))\right] = -E_v\left[\log_2(1)\right] = 0 \text{ bits},$$

$$h(v|u=1) = -E_v\left[\log_2(P(v|u=1))\right] = -E_v\left[\log_2(1)\right] = 0 \text{ bits},$$

$$h(v) = -E_v\left[\log_2(P(v))\right] = -E_v\left[\log_2(0.5)\right] = -\log_2(0.5) = 1 \text{ bit},$$

so that $I(u,v) = 1$ bit. This is the maximum MI that can be achieve between the square wave u and *any* other waveform v. Since only 1 bit of information is encoded in u, only 1 bit of MI can exist between u and *any* v.

3.3 Estimating Entropies

Evaluating equation (2) lies in computing $h(v)$, $h(v|u=0)$, and $h(v|u=1)$. Computing these entropies require $P(v)$, $P(v|u=0)$, and $P(v|u=1)$. In general, we do not have access to these probability densities and hence cannot calculate the entropies directly. We choose a nonparametric method using the leave-one-out procedure to estimate the entropy of an RV from a sample. We employ Parzen window density approximation technique [6] in which windowing functions, centered on the various samples of the RV, are superposed to yield an estimate of the RV's probability density. For convenience, we choose the Gaussian density as the windowing function. To be explicit, our estimate for $P(v)$ is

$$\hat{P}(v) = \frac{1}{(N_{S_v} - 1)}\left[\sum_{v_j \in S_v} G_\sigma(v - v_j) - G_\sigma(0)\right]$$

where N_{S_v} is the number of data points in the sample set S_v and G_σ is the Gaussian density function with σ as the standard deviation of the density function. Set S_v is composed of *all* the data points from v. Our estimate for $P(v|u=0)$ is

$$\hat{P}(v|u=0) = \frac{1}{(N_{S_{v|u=0}} - 1)}\left[\sum_{v_j \in S_{v|u=0}} G_\sigma(v - v_j) - G_\sigma(0)\right]$$

where $N_{S_{v|u=0}}$ is the number of data points in the sample set $S_{v|u=0}$. The sample set is composed of the subset of data points from v with time points corresponding to when $u = 0$. Our estimate for $P(v|u=1)$ is

$$\hat{P}(v|u=1) = \frac{1}{(N_{S_{v|u=1}} - 1)}\left[\sum_{v_j \in S_{v|u=1}} G_\sigma(v - v_j) - G_\sigma(0)\right]$$

where $N_{S_{v|u=1}}$ is the number of data points in the sample set $S_{v|u=1}$. The sample set is composed of the subset of data points from v with time points corresponding to when $u = 1$.

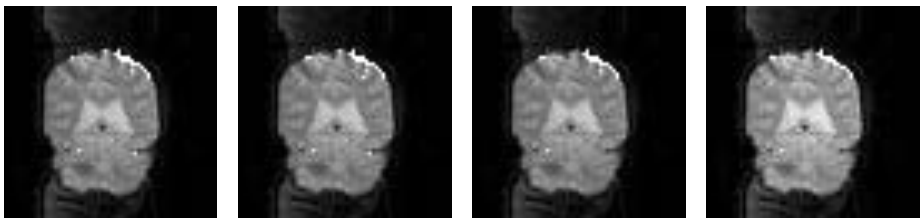
We have found that the Parzen windowing method of estimating the probability density functions is sensitive to the choice of the σ used for the Gaussian windowing function, i.e. the kernel size of the windowing function. We chose the kernel size that maximizes the log-likelihood of the data set used for the density estimation. For example, the kernel size $\hat{\sigma}_{ML}$ used to estimate $h(v)$ is

$$\hat{\sigma}_{ML} = \arg\max_{\sigma} \left\{ \log \frac{1}{(N_{S_v} - 1)} \left[\sum_{v_j \in S_v} G_\sigma(v - v_j) - G_\sigma(0) \right] \right\}.$$

The next step in calculating the entropies is in evaluating the expectations. Direct evaluation of these expectations is difficult so sample mean is used as an approximation to the expectations [8, 9]. Using this approach, we obtain approximations to the entropies $h(v)$, $h(v|u = 0)$, and $h(v|u = 1)$.

4 Experimental Results

We applied the above described fMRI analysis method to a set of fMRI data that examines right-hand movements. The data set contains 60 whole brain acquisitions with each whole brain acquisition containing 21 slice images.



(a) DS (b) CC (c) GLM (d) MI

Fig. 2. Comparison of fMRI Analysis Techniques.

Only the analysis results from the 10th coronal slice of the whole brain acquisition are shown in Fig. 2. The figure provides a qualitative comparison of our analysis technique with other techniques previously mentioned in this paper. A quantitative comparison of these different methods is difficult since the ground truth is unknown. In keeping with the fairness of the comparison, the threshold (which determines whether a voxel is activated or not) that yields the "best" activation map for each analysis technique is used. For this particular fMRI data set, the "best" activation map is judged based on the prior expectation that brain activation is restricted to the left primary motor cortex and occurs in clusters. It is important to point out that MI is inherently a normalized measure so for our technique, the threshold can be specified meaningfully in terms of bits

of information. Fig. 2(d) is obtained using a threshold of 0.7 bits. It can be seen that all four techniques yield similar results. As mentioned earlier and as evident in Fig. 2, it is difficult to determine whether one method is better than another. We can, however, conclude from the experimental results that MI can offer a new viable alternative for fMRI data analysis.

5 Summary

We have developed a theoretical framework for using MI to calculate the fMRI activation map. While there are many existing approaches to calculate the activation map, all these techniques depend on some *a priori* assumptions about the relationship between the protocol timeline and the fMRI voxel temporal response. The strength of our approach is that it relies on sound theoretical principles, it is fairly easy to implement, and does not require strong assumptions about the nature of the relationships between the fMRI temporal measurements and the protocol timeline, while still retaining the ability to uncover complex relationships (beyond second-order statistics). In addition, experimental results confirmed that this information-theoretic approach can be as effective as other methods of calculating activation maps.

References

1. P.A. Bandettini, A. Jesmanowicz, E.C. Wong, and J.S. Hyde. Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, 30:161–173, 1993.
2. F. Bello and A.C.F. Colchester. Measuring global and local spatial correspondence using information theory. In *Proceedings of the First International Conference on Medical Computing and Computer-Assisted Intervention*, 1998.
3. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Son Inc., 1st edition, 1991.
4. K.J. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1:153–171, 1994.
5. O. Henriksen, H.B.W. Larsson, P. Ring, E. Rostrup, A. Stensgaard, M. Stubgaard, F Stahlberg, L. Sondergaard, C. Thomsen, and P. Toft. Functional MR imaging at 1.5T. *Acta Radiologica*, 34:101–103, 1993.
6. E. Parzen. On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.
7. D.L. Schacter and R.L. Bucknew. On the relations among priming, conscious recollection,and intentional retrieval: evidence from neuroimaging research. *Neurobiology of Learning and Memory*, 70(1):284–303, 1998.
8. P. Viola and W.M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.
9. W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.
10. G.K. Wood, B.A. Berkowitz, and C.A. Wilson. Visualization of subtle contrast-related intensity changes using temporal correlation. *Magnetic Resonance Imaging*, 12(7):1013–1020, 1994.