Adaptive Entropy Rates for *f*MRI Time-Series Analysis*

John W. Fisher III[1], Eric R. Cosman, Jr.[1], Cindy Wible[2], and William M. Wells III[1,2]

[1] Massachusetts Institute of Technology,
Artificial Intelligence Laboratory,
Cambridge, MA, USA
{fisher, ercosman, sw,}@ai.mit.edu
[2] Harvard Medical School,
Brigham and Women's Hospital,
Department of Radiology,
Boston, MA, USA
cindy@bwh.harvard.edu

Abstract. In previous work [Tsai *et al* , 1999] we introduced an information theoretic approach for analysis of *f*MRI time-series data. Subsequently, [Kim *et al* , 2000] we established a relationship between our information theoretic approach and a simple non-parametric hypothesis test. In this work, we describe an adaptive approach for incorporating the temporal structure that relates the *f*MRI time-series to both the current *and* past values of the experimental protocol. This is achieved via an extension of our previous approach using the information-theoretic concept of entropy rate. It can be shown that, despite a differing implementation, our prior method is a special case of the new approach. The entropy rate of a random process quantifies future uncertainty conditioned on the past and side-information (e.g. the experimental protocol, confounding signals, etc.) without making strong assumptions about the nature of that uncertainty (e.g. Gaussianity). Furthermore, we allow the form of the dependency to vary from voxel to voxel in an adaptive fashion. The combination of the information theoretic principles and adaptive estimation of the temporal dependency allows for a more powerful and flexible approach to *f*MRI analysis. Empirical results are presented on three *f*MRI datasets measuring motor, auditory, and visual cortex activation comparing the new approach to the previous one as well as a variation on the general linear model. Particular attention is paid to the differences in the type of phenomenology detected by the respective approaches.

* J. Fisher was supported in part by an NSF ERC grant, Johns Hopkins Agreement #8810274. E. Cosman was supported under the NSF award #IIS-9610249. C. Wible was supported in part under NIMH grants MH40799 and MH52807. W. Wells was supported by the same ERC grant, and by NIH 1P41RR13218.

W. Niessen and M. Viergever (Eds.): MICCAI 2001, LNCS 2208, pp. 905–912, 2001.

1 Introduction

Previously, we have discussed the application of an information theoretic formalism to the analysis of *f*MRI time series. In [8] we presented a novel information theoretic approach for calculating *f*MRI activation maps by estimating the mutual information between an encoding of the experimental protocol and *f*MRI voxel time-series. Subsequently in [5] we demonstrated the equivalence of the method to a statistical hypothesis test when the underlying densities are unknown. As a consequence, the computation of the activation map can be formulated as a binary MAP detection problem using the Ising model as a spatial prior and solved *exactly* in polynomial time using the Ford and Fulkerson method [4]. The information-theoretic framework is appealing in that it is a principled methodology requiring few assumptions about the structure of the *f*MRI signal. It is capable of detecting unknown nonlinear and higher-order statistical dependencies. Furthermore, it is relatively straightforward to implement. An implicit assumption of [8] and [5] is that samples of the time-series are statistically independent. That is, time structure was neither assumed nor exploited. This is in contrast to approaches based upon the general linear model (GLM) in which strong assumptions about the time structure are made through the choice of a set of basis vectors or equivalently a signal subspace [3].

In this work, we consider a natural extension to the information theoretic method in which we learn and then exploit the time structure of the *f*MRI voxel time-series and its dependence on the time structure of the protocol. Whereas when we assumed sample independence, mutual information was a natural way to relate the protocol to the *f*MRI time series, the information theoretic notion of *entropy rate* is the natural quantity when we consider time structure. As a consequence of the adaptive learning approach, we do not make strong assumptions about the exact nature of the time structure *á priori*, merely that it exists and can be estimated. In fact, time dependence of the *f*MRI time-series is allowed to vary from voxel to voxel. In doing so, the determination of whether or not a voxel is declared active relies upon the methodology described in [2]. That work discusses a general approach for random process analysis. The assumption is that the dependency is distributed across many samples in the past, but may be approximated using low-dimensional functions of the past. In this work we examine a special case for which the methodology of [2] is appropriate.

We compare the new method to an approach based on the general linear model (GLM) popularized by Friston *et al* [3] using data from three *f*MRI data sets testing motor, auditory, and visual cortex activation.

2 Entropy Rates

We model the *f*MRI time series as a random process, denoted $\{Y\} \equiv \{y\}_0^\infty \equiv \{y_0, y_1, \cdots\}$ in which a sample y_k statistically depends on the past values $\{y\}_0^{k-1} \equiv \{y_0, \cdots, y_{k-1}\}$, and perhaps on the present and past values of the protocol time-series, $\{u\}_0^k$. This dependence is quantified by the information theoretic notion of *entropy rate*, defined as [1]

$$H(Y) \equiv \lim_{N\to\infty} \frac{1}{N+1} h\left(\{y\}_0^N\right) = \lim_{N\to\infty} h\left(y_N | \{y\}_0^{N-1}\right) \tag{1}$$

where $h(\ )$ is differential entropy. Note that equality assumes the process is stationary, however, the second form is also valid for a wide class of nonstationary processes and is the form we use in practice. The entropy rate quantifies the average uncertainty about future values conditioned on the past. We can also condition on side information (e.g. the protocol) by

$$H(Y|\{u\}_0^N) = \lim_{N\to\infty} h\left(y_N | \{y\}_0^{N-1}, \{u\}_0^N\right) \tag{2}$$

It can be shown that, in general, conditioning reduces entropy, that is

$$H(Y|\{u\}_0^N) \leq H(Y) \tag{3}$$

with equality only when $\{y\}_0^N$ and $\{u\}_0^N$ are statistically independent [1]. Equations 1 and 2 imply that we must consider the joint densities over all samples of the process which is generally intractable. However, we make two assumptions that reduce the complexity. First, we assume the process depends on the finite past, that is

$$h\left(y_k | \{y\}_0^{k-1}\right) = h\left(y_k | \{y\}_{k-M_y}^{k-1}\right) \tag{4}$$

$$h\left(y_k | \{y\}_0^{k-1}, \{u\}_0^k\right) = h\left(y_k | \{y\}_{k-M_y}^{k-1}, \{u\}_{k-M_u}^k\right) \tag{5}$$

limiting the dimensionality to $M_y + M_u + 1$. Furthermore, we assume the information about y_k in the samples $\{y\}_{k-1}^{k-M_y}$ and $\{u\}_k^{k-M_u}$ can be summarized by lower dimensional functions,

$$h\left(y_k | \{y\}_{k-M_y}^{k-1}\right) \approx h\left(y_k | f_a\left(\{y\}_{k-M_y}^{k-1}\right)\right) \tag{6}$$

$$h\left(y_k | \{y\}_{k-M_y}^{k-1}, \{u\}_{k-M_u}^k\right) \approx h\left(y_k | f_a\left(\{y\}_{k-M_y}^{k-1}\right), f_b\left(\{u\}_{k-M_u}^k\right)\right) \tag{7}$$

where $f_a(\ )$ and $f_b(\ )$ are parameterized. Fisher *et al* [2] describe a generalized approach for both learning the parameterized functions and then using them to compute entropy rates. Note that when using the methodology of [2] the approximation of equations 6 and 7 are close in the Kullback-Leibler sense and thus consistent with a hypothesis testing framework [6]. For reasons of brevity we shall only consider the class of linear predictive models.

2.1 Hypothesis Testing and Entropy Rates

For the moment we put aside the question of estimating the functional parameters to examine the relationship between hypothesis testing and entropy rates. Consider the following hypothesis test.

$$H_0 : y_k \sim p_Y(Y_k | \{y\}_0^{N-1})$$
$$H_1 : y_k \sim p_{Y|U}(Y_k | \{y\}_0^{N-1}, \{u\}_0^N)$$

Hypothesis H_0 states that the random process Y depends only on the past of Y, while H_1 states that the random process depends on the past of both Y and U. We compute the log of the likelihood ratio

$$T_n = \sum_{k=1}^{n} \log\left(p_{Y|U}(y_k|\,\{y\}_0^{N-1}, \{u\}_0^{N})\right) - \log\left(p_Y(y_k|\,\{y\}_0^{N-1})\right) \ . \tag{8}$$

It can be shown that [1]

$$\lim_{n\to\infty} T_n = n\left(H\left(Y\right) - H\left(Y|\,\{u\}_0^{N}\right)\right) = \mathrm{E}\left\{T_n\right\} \ , \tag{9}$$

consequently, the difference in entropy rates used as the activation statistic is equivalent to the aforementioned hypothesis test. In practice we substitute the Parzen density estimate and, as in equations 6 and 7, substitute functions $f_a(\ )$ and $f_b(\ )$ to summarize the dependency on the past.

The preceding analysis is similar to that presented in [5] which shows the equivalence between a simpler hypothesis test and mutual information as the test statistic. We note here that when the process is independent from sample to sample and is dependent on the coincident (or a delayed) sample of the protocol then the hypothesis test is equivalent to that described in [5]. In [8], the test was performed for a range of delays modeling a type of hemodynamic response. Here, since we consider past samples jointly this is not necessary and consequently more general forms of the hemodynamic response are modeled. This will be further elaborated in the experimental results section.

3 Modeling the Temporal Structure of *f*MRI Time-Series

We now discuss a method for estimating the time dependence of the *f*MRI time-series y_k on past samples of both y_k and the protocol u_k. We restrict ourselves to linear functionals of past samples, a special case of the more general approach described in [2]. Note the information theoretic principles play a role both in quantifying the dependence of the voxel time-series on the protocol *and* in estimating the parameters of the functions. Letting y_k represent the time-series at some voxel, we consider two signal components.

$$y_k^a = -\sum_{i=1}^{M_y} a_i y_{k-i}^a + n_k \qquad y_k^b = \sum_{i=0}^{M_u} b_i u_{k-i} \tag{10}$$

where n_k is an i.i.d. noise sequence and u_k is the protocol signal. We assume that n_k is independent of u_k. The estimates of $\{a_i\}$ and $\{b_i\}$ should reflect this assumption. This condition partially distinguishes our approach from standard ARMA (auto-regressive moving-average) models. Consequently, y_k^a is an AR (auto-regressive) process with a random noise source while y_k^b is a MA (moving-average) process with the protocol as the input. The *f*MRI time-series is modeled as the sum of these two signals.

$$y_k = y_k^a + y_k^b = y_k^a + \sum_{i=0}^{M_u} b_i u_{k-i} = \sum_{i=1}^{M_y} a_i y_{k-i}^a + n_k + \sum_{i=0}^{M_u} b_i u_{k-i} \tag{11}$$

Ordinarily, solving for ARMA parameters is a nonlinear optimzation problem. However, since u_k is known, the parameters can be obtained using linear least squares estimation [7] with the additional constraint that n_k is statistically independent of u_k. Furthermore, it can be shown for linear predictive models (such as in equation 10) that the solution obtained by minimizing the squared error in the prediction has the same expected value as that obtained by minimizing the entropy rate (equations 6 and 7).

Standard least squares methods do not ensure the independence of n_k and u_k, in general. If an active voxel obeys the model (even approximately) past samples will contain some dependence on the protocol which will be transmitted through the $\{a_i\}$ coefficients. That dependence must be removed so that we can isolate contributions from u_k and pass them only through the $\{b_i\}$ coefficients. A straigtforward approximation is to solve for the coefficients sequentially.

$$\{b_i\} = \arg\min_{\{b_i'\}} \sum_k \left(y_k - \sum_{i=0}^{M_u} b_i' u_{k-i} \right)^2 \tag{12}$$

$$\{a_i\} = \arg\min_{\{a_i'\}} \sum_k \left(y_k - y_k^b - \sum_{i=1}^{M_y} a_i' y_{k-i}^a \right)^2 \tag{13}$$

To the degree that the model order is correct, predictions of y_k from y_k^a will be independent of y_k^b. Furthermore, the entropy rates of the processes are equivalent to the entropy of the error residuals (with and without the protocol contribution). That is, under the model,

$$H(Y) = h(y_k - y_k^a) \quad \text{and} \quad H(Y|\,\{u\}_0^k) = h(y_k - y_k^a - y_k^b), \tag{14}$$

where $(y_k - y_k^a)$ and $(y_k - y_k^a - y_k^b)$ are our estimates of n_k under the two hypotheses. As follows from equation 8, the difference of these entropies form our ARMA-based entropy rate statistic (*ER*).

Note that the MA model y_k^b is the same model implicit in a GLM approach when the basis is $M_u + 1$ shifted versions of the protocol signal. This is only the case when we restrict ourselves to *linear* functions of the past, and represents a principled way to choose a GLM design matrix and to include noise modeling. In Section 4, we contrast the ER-test to a classical F-test that approximates incorporating our modified ARMA model into a GLM design matrix.

4 Empirical Results and Discussion

We present results on three fMRI datasets, whose respective protocols were designed to activate the motor cortex (dominant hand movement protocol), auditory cortex (verb generation protocol), and visual cortex (visual stimulation

with alternating checkerboard pattern). Each data set contains 60 whole brain acquisitions taken three seconds apart, each consisting of 21 coronal slices. Each protocol consists of a 30 second rest phase followed by a 30 second task phase repeated three times. In all cases, the MA and AR systems are sixth-order.

We compare the two versions of the new method to a baseline GLM (the protocol signal is the basis) and MI [8]. The first estimates the entropy rates nonparametrically (ER) while the second assumes Gaussian statistics, whereby the difference in entropy rates is equivalent to a variance ratio (an F-test, and in effect, a GLM with our ARMA model in the design matrix). For each protocol the GLM threshold was set so that the number and location of activations was consistent with the protocol. In the comparison methods, the thresholds were set such that all of the GLM activations were detected. This resulted in additional detections, some spurious and some not. Results for the visual cortex are shown in Figure 1. Visual inspection of the signals of the new activations in the ER maps (contrast Figure 1 (a) and (c)) suggested a relationship to the protocol. Figures 2 and 3 present four such activations, from visual, auditory, and motor protocols, which the GLM only detects when its threshold is siginificantly lowered. The lowered GLM threshold produced many additional detections as well. This is illustrated in Figure 4, which contrasts the detections in slice 6 before and after the threshold change. In this particular case, all activations detected with the lower threshold (Figure 4(d)) had very small MI and ER values and were judged to be spurious by inspection. One such false positive is shown in Figure 5.

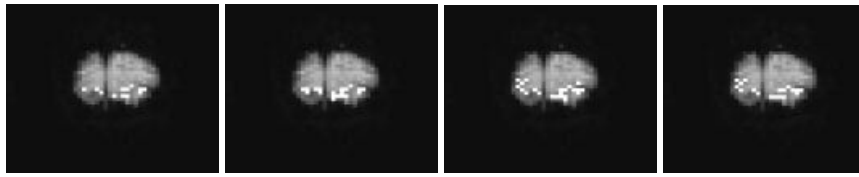
We found that the ARMA-based F-statistic and the GLM statistic are both small for the active voxels plotted in Figure 2 (a) and (e). This is likely due to their inherent assumption of Gaussianity. Figure 2 (c) and (g) suggests a bi-modal error density for these cases. The ER statistic does not make this assumption and can more robustly detect these depedencies on the protocol.

The hemodynamic response is modeled implicitly by the MA term in the model. When calculating MI and GLM statistics, this delay is found by searching a range of delays. Figure 2(b) shows the MA signal estimate as a solid line overlying the dotted line representing the protocol period (and the *f*MRI signal, dashed) and is slightly delayed. Note that this delay corresponds roughly to the centroid or peak of the MA weights shown in Figure 2(d), which is between 2 and 3 delays. The active voxel shown in Figure 3 is one whose MI value erroneously indicates a vanishing dependency on the protocol, but which ER accurately detects. We have observed that MI and GLM poorly detect signals like this one, which are characterized by high-amplitude, high-frequency variation (relative to the protocol period).

While these results are not exhaustive, we feel that they are indicative of the potential of the method, particularly in the cases where it is difficult to model phenomenon *á priori*.

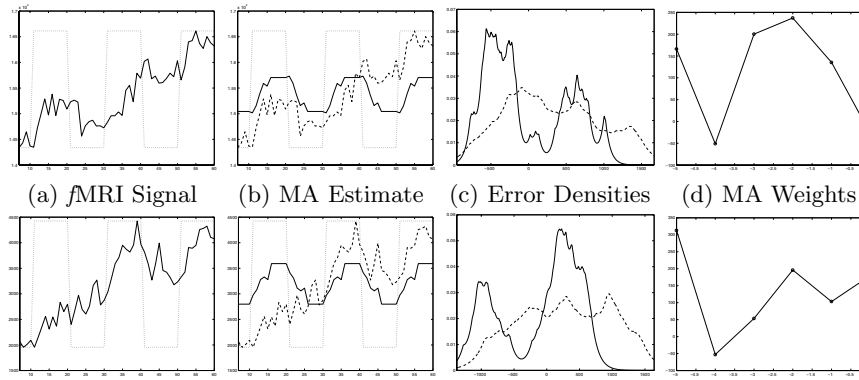
References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, 1991.



(a) GLM (b) MI (c) ARMA ER (d) ARMA F-test

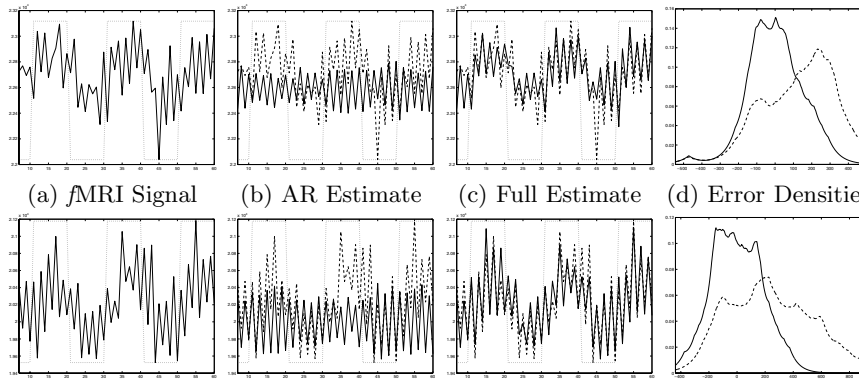
Fig. 1. Comparison of *f*MRI analysis results from visual experiments (2nd slice)



(a) *f*MRI Signal (b) MA Estimate (c) Error Densities (d) MA Weights

(e) *f*MRI Signal (f) MA Estimate (g) Error Densities (h) MA Weights

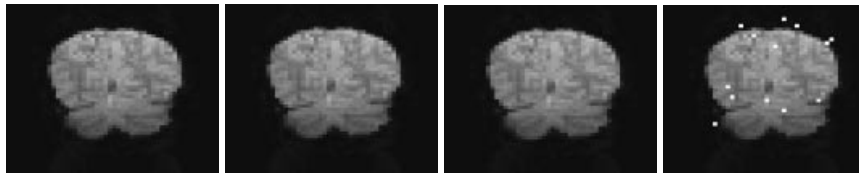
Fig. 2. The "partial-responders" (a,e) appear to be cases in which the subject did not respond during all task phases (visual - top, auditory - bottom). Consequently, the error residuals exhibit a bimodal density (solid lines in (c,g)). The residual error density using only past signal values (dotted lines) has lower entropy, but similar variance, so ER detects the lower entropy rate, while ARMA F-test and baseline GLM do not.



(a) *f*MRI Signal (b) AR Estimate (c) Full Estimate (d) Error Densities

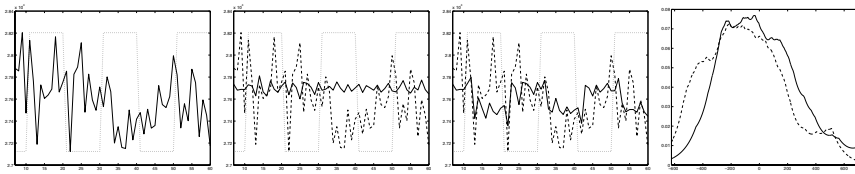
(e) *f*MRI Signal (f) AR Estimate (g) Full Estimate (h) Error Densities

Fig. 3. The ER statistic finds dependencies on the protocol that was apparently hidden from both MI and GLM by the low SNR in both these signals. (a)-(d): A voxel in the visual protocol (e)-(h): A voxel in the motor protocol



(a) GLM (original) (b) MI (c) ARMA ER (d) GLM (lower)

Fig. 4. Visual experiment, slice 6: spurious GLM detections due to lowered threshold which do not occur in ER result.



(a) *f*MRI Signal (b) AR Estimate (c) Full Estimate (d) Error Densities

Fig. 5. A voxel from the 6th *f*MRI slice during a visual experiment, erroneously detected by the GLM with a lowered threshold, but with a vanishing MI or ER value

[2] J. W. Fisher III, A. T. Ihler, and P. A. Viola. Learning Informative Statistics: A Nonparametric Approach. In S. A. Solla, T. K. Leen, and K. R. Müller, editors, *Proceedings of 1999 Conference on Advances in Neural Information Processing Systems 12*, 1999.

[3] K. J. Friston, P. Jezzard, and R. Turner. The Analysis of Functional MRI Time-series. *Human Brain Mapping*, 1:153–171, 1994.

[4] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B(Methodological)*, 51(2):271–279, 1989.

[5] J. Kim, J. W. Fisher III, A. Tsai, C. Wible, A. Willsky, and W. M. Wells III. Incorporating Spatial Priors into an Information Theoretic Approach for *f*MRI Data Analysis. In S. L. Delp, A. M. DiGioia, and B. L. Jaramaz, editors, *Third International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 1935 of *Lecture Notes in Computer Science*, pages 62–71 Springer, Oct 2000.

[6] S. Kullback. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

[7] L. L. Scharf. *Statistical signal process: detection, estimation, and time series analysis*. Addison-Wesley Publishing Company, New York, 1990.

[8] A. Tsai, J. Fisher III, C. Wible, W. M. Wells III, J. Kim, and A. Willsky. Analysis of Functional MRI Data using Mutual Information. In C. Taylor and A. Colchester, editors, *Second International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 1679 of *Lecture Notes in Computer Science*, pages 473–480. Springer, Sept 1999.