The Effect of Correlation and Error Rate Specification on Thresholding Methods in fMRI Analysis

Brent R. Logan[1] and Daniel B. Rowe[2,*]
Division of Biostatistics[1] and Department of Biophysics[2]

Division of Biostatistics
Medical College of Wisconsin

Technical Report 42

May 2003

Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI 53226
Phone:(414) 456-8280

The Effect of Correlation and Error Rate Specification on Thresholding Methods in fMRI Analysis

Brent R. Logan[1] and Daniel B. Rowe[2,*]

Division of Biostatistics[1] and Department of Biophysics[2]

Medical College of Wisconsin

Milwaukee, WI USA

Abstract

This paper reviews and compares thresholding methods for identifying active voxels in single-subject fMRI datasets. Different error rates are described which may be used to calibrate activation thresholds. We discuss methods which control each of the error rates at a pre-specified level α, including simple procedures which ignore spatial correlation among the test statistics as well as more elaborate ones which incorporate this correlation information. The operating characteristics of the methods are shown through a simulation study, indicating that the error rate used has an important impact on the sensitivity of the thresholding method, but that accounting for correlation has little impact for individual voxelwise thresholding methods. The methods are illustrated with a real bilateral finger tapping experiment.

Some Keywords: Familywise Error rate; False Discovery Rate; Spatial Correlation; Permutation Resampling.

$*$ Address correspondence to Daniel B. Rowe, Department of Biophysics, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226-0509, dbrowe@mcw.edu

1. Introduction

Many fMRI experiments have a common objective of identifying active voxels in a neuroimaging dataset. This is done in single subject experiments for example by performing individual voxelwise tests of the null hypothesis that the observed time course is not significantly related to an assigned reference function (Bandettini et al., 1993; Cox et al., 1995). A voxel activation map is then constructed by applying a thresholding rule to the resulting t-statistics.

This paper describes three error rates which may be used to formally set activation thresholds based on individual voxelwise test statistics, but not on cluster size. We review methods which control each of the error rates at a pre-specified level α. These methods include simple procedures which ignore spatial correlation among the test statistics as well as more elaborate ones which incorporate this correlation information. The operating characteristics of the methods are shown through a simulation study. A real bilateral finger tapping experiment is used to illustrate the methods and conclusions.

2. Problem and Error Rates

A common way of determining significance of a statistical hypothesis test is to specify the significance level or type I error rate of the test, commonly denoted by α, and use this to determine a threshold. The type I error rate is the probability that, if the voxel were truly inactive, its test statistic would exceed the threshold and we would incorrectly conclude that it is active. This significance level determines the threshold, so that for example, a 5% level voxel z-test would have a threshold of 1.96 (two-sided) or 1.645 (one-sided). However, there is an important problem here. If we consider for example a $64 \times 64 \times 15$ volume image, as in the real experiment which follows, with no true activity attributable to treatment, we would expect $.05 \times 61440 = 3072$ voxels to exceed a 5% threshold by chance alone. Therefore, when we use this kind of thresholding rule, the result is a large number of false positives, or voxels declared active when they are truly inactive. The reason for this problem is that there are multiple individual voxel hypotheses being tested (called the multiplicity problem).

As a result of this problem of excessive false positives, it is useful to consider other types

of error rates which account for the multiplicity problem. Some notation needs to be laid out before proceeding. For voxels $i = 1, \ldots, m$, define t_i to be the test statistic for treatment-related activation at voxel i, T_i to be the random variable corresponding to the observed t_i, $p_i = P(|T_i| > |t_i|)$ to be the (two-sided) p-value for voxel i, μ_i to be the actual treatment-related activation for voxel i, and γ to be the fixed threshold set for determining whether a voxel is active. For example, $\mu_i = 0$ if voxel i is truly inactive and $\mu_i \neq 0$ if voxel i is truly active. We define $E_i : \{|t_i| > \gamma | \mu_i = 0\}$ to be the event that the test statistic in voxel i exceeds the threshold γ when voxel i is truly inactive.

When no account is made of the multiple testing, the error rate is the usual significance level or type I error rate. We will also call it the per comparison error rate (or per voxel error rate) and it refers to the probability of a false positive finding for an individual voxel i. Using our notation, the per comparison error rate (PCE) is the probability of the event E_i,

$$PCE = P(E_i).$$

The most common way in the statistical literature to account for multiplicity is to consider the familywise error rate (or imagewise error rate). The familywise error rate (FWE) is the probability of at least one false positive on any voxel in the image,

$$FWE = P(\cup_i E_i).$$

Note that $FWE \geq PCE$ so that any method which controls the FWE at level α will have a higher threshold γ than one which controls the PCE at level α. Sometimes a distinction is made between methods which only control the FWE under the overall null hypothesis (no voxels have treatment-related activity), called weak control of the FWE, and those which control the FWE under any null hypothesis (any subset of voxels have no treatment-related activity), called strong control of the FWE.

Recently there has been much interest in a third criterion called the false discovery rate (FDR). The FDR is the expected proportion of false positives to total positives, or the expected proportion of truly inactive voxels which are declared active to the total number of voxels declared active. This is illustrated in Table 1, where the letters in each cell refer to the counts of the number of individual hypotheses falling in the corresponding category. For example, V is the number of inactive voxels which are declared to be active.

Table 1: True Status vs. Decision for all m voxels

	Decision		
True Status	Declared Inactive	Declared Active	Total
voxel inactive	U	V	m_0
voxel active	T	S	m_1
	Y	R	m

Then the false discovery rate is

$$FDR = E(V/R),$$

where the ratio V/R is defined to be 0 when $R = 0$. Note that using this notation,

$$FWE = P(V > 0).$$

In the case where all null hypotheses are true (called the global null hypothesis), then the number of false positives V is equal to R, so that $V/R = 1$ if $R > 0$ and 0 otherwise. The FDR under this scenario simplifies to

$$FDR = P(R > 0) = P(V > 0) = FWE.$$

Therefore, any FDR-controlling procedure can be said to have weak control of the FWE.

3. Methods for Controlling the FWE

The simplest way to control the FWE is through the Bonferroni method. To apply this, simply divide the individual threshold significance level α by the number of voxel hypotheses m to arrive at an adjusted threshold significance level $\alpha' = \alpha/m$ for each voxel test. This guarantees that the FWE is no larger than α, since

$$FWE = P(\cup_i E_i) \leq P(E_1) + \cdots + P(E_m) = m\alpha' = \alpha.$$

One limitation of the Bonferroni method is that it results in conservative control of the FWE (i.e. fewer voxels declared active) in many situations because the approximation

ignores intersections of events E_i. This conservativism is usually most severe when the test statistics are moderately to strongly correlated. Functional MRI data is known to exhibit spatial autocorrelation, where closely spaced voxels are more strongly correlated with one another. The result of the conservative behavior of the Bonferroni method is potentially less power to detect truly active voxels.

To sharpen the Bonferroni procedure and obtain less conservative control of the FWE (i.e. more voxels declared active), one must set thresholds based on the distribution of the maximum $|T|$ statistic. This is because the FWE for threshold γ can be written as

$$FWE = 1 - P(|T_1| \leq \gamma, \ldots, |T_m| \leq \gamma) = 1 - P(\max |T_i| \leq \gamma),$$

so that the exact threshold γ to obtain a FWE of α is the $(1-\alpha)$ percentile of the maximum $|T|$ distribution. Equivalently, one can consider thresholds based on the minimum voxel p-value. This distribution is dependent on the correlation structure of the t statistics, and may be obtained in several ways.

Random field methods were first applied to functional neuroimaging data to approximate this max $|T|$ distribution by Friston et al. (1991) and Worsley et al. (1992). They assume that the m t-statistics can be viewed as a lattice representation of a continuous Gaussian random field of volume S voxels and smoothness parameter W. This representation is then used to estimate the exact γ threshold value corresponding to $FWE = \alpha$, i.e. the $(1-\alpha)$ percentile of the max $|T|$ distribution. Equivalently we can estimate an adjusted p-value for p_i, which is the probability that the maximum $|T|$ statistic exceeds $|t_i|$ (or equivalently the probability that the minimum p-value is smaller than p_i). This adjusted p-value can be compared directly to α to determine if voxel i is active. Using this Gaussian random field theory, Friston et al. (1991) approximate the adjusted p-value for p_i by

$$P_{RF}(p_i) \approx 1 - \exp(-E(d)),$$

where $E(d)$ is the expected number of maxima or clusters of voxels exceeding $|t_i|$, computed by

$$E(d) \approx S(2\pi)^{-(m+1)/2} W^{-m} p_i^{-m} \exp(-p_i^2/2). \tag{3.1}$$

Worsley et al. (1992) elaborate on this approximation by using the expected Euler characteristic in place of $E(d)$. Alternatively, these Gaussian random fields may be simulated.

Further work on random field theory is reviewed in Petersson et al. (1999)

Holmes et al. (1996) present an alternative way of deriving the distribution of the maximum $|T|$ statistic. They propose to simulate the exact γ values using permutation resampling of the multiple scans over time. If under the null hypothesis the data from these scans are exchangeable (have the same distribution), we can generate the exact empirical distribution of the max $|T|$ statistic by enumerating each permutation, recomputing each voxel t statistic, and determining the observed max $|t|$ statistic across voxels for each permutation. In practice, one takes a random sample of possible permutations rather than enumerating each one, because the number of permutations becomes prohibitively large with the number of time points. For example, with $n = 128$ time points, there would be $128! = 3.8 \times 10^{215}$ possible permutations. A random sample of these permutations yields a max T distribution which converges to the true distribution with increasing sample size.

To obtain exchangeability, two modifications may need to be made to the data. A time trend may need to be accounted for, so that one instead would permute the residuals after fitting a model for the time trend. Also, the time courses may exhibit temporal autocorrelation. Locascio et al. (1997) estimated the temporal autocorrelation using a parametric model, whitened the data based on this estimate, and then applied the permutation procedure to the whitened data in order to estimate the distribution of the maximum $|T|$ statistic. Fitting a time trend and whitening the data make the residuals approximately exchangeable, because the parameters computed from the data are only estimates of the true values. In practice, this type of procedure works well assuming the model is correct.

In conclusion, methods for controlling the FWE require a threshold to be set based on the distribution of the maximum $|T|$ statistic. Many possibilities for estimating this distribution exist, but the permutation resampling method is especially attractive because it can be applied to many situations, provided one can fit an appropriate model so that the residuals are exchangeable.

While several methods exist for controlling the FWE in fMRI data, it is important to consider whether the FWE is a relevant criterion for fMRI data. Is it relevant to focus on the probability of getting one or more false positives in a volume image of $61,440$ voxels? The consequence of controlling the FWE at a set level α means that we will have relatively

low power to detect truly active voxels. As in the Bonferroni procedure, we are adjusting the threshold for such a large number of voxel tests, and the signal or activity level will have to be very strong to be above this adjusted threshold.

One way of mitigating this problem is to consider a priori defined regions of interest (ROI). One can control the FWE in a region of the image which is of specific interest, using Bonferroni adjustment or some other FWE method. This has the advantage that there is less multiplicity adjustment because of reduced family size (reduced number of voxels). Therefore, this method will have higher power to detect active voxels in that region. The disadvantage is that these ROI's must be identified a priori and must remain unchanged throughout the experiment and analysis. Otherwise, the FWE over that region will no longer be controlled. One example of an a priori identified ROI is to apply a mask to the image, so that only voxels inside the brain are considered in the multiplicity adjustment. This reduces the total number of voxel hypotheses and improves the power to detect activated voxels inside the brain and uncovered by the mask. Also, it is not difficult to specify a mask a priori, so this is a straightforward way to reduce the multiplicity adjustment. For simplicity, however, we do not consider masking further in the remainder of this paper.

The FWE criterion considers it unacceptable to have a false positive occurring anywhere in the thresholded volume image. An alternative is to allow some false positives in the thresholded image, but to relate the number of these acceptable false positives to the number of total positive findings. This is the basic strategy of controlling the FDR, and will be discussed next.

4. Methods for Controlling the FDR

Benjamini and Hochberg (BH, 1995) propose a simple step-up procedure for controlling the FDR at level q, which was applied to neuroimaging data by Genovese et al. (2002). First order the voxel p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Let $v_{(i)}$ denote the voxel corresponding to p-value $p_{(i)}$. Then let r be the largest i for which

$$p_{(i)} \leq \frac{i}{m} q.$$

The BH procedure declares voxels $v_{(1)}, \ldots, v_{(r)}$ to be active. It is called a step-up procedure because of the sequential or stepping up method for finding r.

This procedure controls the FDR at level q for independent and positively dependent voxel test statistics (Benjamini and Yekutieli, 2001). For a general correlation structure with potential negative correlations, Benjamini and Yekutieli (2001) show that the FDR is still controlled if you redefine r above to be the largest i for which

$$p_{(i)} \leq \frac{i}{m} \frac{q}{\sum_i 1/i}.$$

However, this will result in smaller r and fewer voxels declared active. So it is preferable to use the first procedure unless negative correlations are likely.

Storey (2001a) uses a simplified version of the FDR, which he calls the positive FDR (pFDR):

$$pFDR = E(V/R|R > 0).$$

It turns out that under independence of the test statistics, the pFDR has a natural Bayesian interpretation (Storey 2001b) as

$$pFDR = p = P(H_i \text{ is true}||T_i| > \gamma),$$

or the posterior probability that the voxel is inactive, given that its test statistic is above the threshold. To see this, note that conditional on $V + S = R$, then V, the number of voxels which are falsely declared active is Binomial($V + S$,p) where p is the probability that H_i is true given that $|T_i| > \gamma$. Therefore,

$$\begin{aligned} pFDR &= E(V/R|R > 0) \\ &= E_{R>0}[E(V/R|R)] \\ &= E_{R>0}[p(V + S)/(V + S)] = p. \end{aligned}$$

Storey (2001a,b) as well as Storey and Tibshirani (2001) use the Bayesian interpretation to express the pFDR in the following way:

$$\begin{aligned} pFDR &= \frac{P(|T_i| \geq \gamma|\mu_i = 0)P(\mu_i = 0)}{P(|T_i| \geq \gamma)} \\ &= \frac{F_0(\gamma)\pi_0}{F(\gamma)}, \end{aligned}$$

where F_0 is the complement of the CDF of $|T_i|$ when voxel i is inactive, F is the complement of the marginal CDF of $|T_i|$ regardless of activity (a mixture distribution of F_0 and an

unknown alternative distribution F_1), and $\pi_0 = m_0/m$ is the proportion of null hypotheses (or inactive voxels).

They discuss both parametric and nonparametric ways of estimating each of these quantities. Parametric models assume that p-values can be reliably computed from the observed voxel test statistics, while nonparametric methods use resampling to estimate the distribution of the voxel test statistics. Parametric estimation turns out to be identical to the BH method. Suppose that the p-value $p_{(i)}$ corresponded to the threshold $\gamma_{(i)}$ i.e. $p_{(i)} = F_0(\gamma_{(i)})$. We can estimate $\hat{F}(\gamma_{(i)})$ by the proportion of rejected hypotheses, so that $\hat{F}(\gamma_{(i)}) = i/m$. A worst case or conservative estimate of pFDR can be obtained by using $\pi_0 = 1$. Then $pFDR = mp_{(i)}/i$ so that

$$pFDR \leq q \iff p_{(i)} \leq \frac{i}{m} q.$$

Therefore this parametric Bayesian formulation relating the p-values to their corresponding t-statistics turns out to be equivalent to the BH method.

Nonparametric estimates of F_0, which could be obtained using either permutation resampling or Gaussian random field methods discussed above, could also be incorporated into the pFDR estimate. However, as these are meant to simulate the marginal null distribution, and not to incorporate the correlation information into the FDR estimate, we do not discuss them further.

Several authors (Storey, 2001a,b; Benjamini and Hochberg, 2000) have considered adaptive estimation of π_0 to further refine the FDR-controlling procedure. However, in most fMRI datasets, we expect that a relatively small proportion of the voxels in an image would actually be considered active as a result of treatment. In this setting, there may be limited utility in estimating π_0 because the estimate will typically be very close to one.

The BH method is a simple, powerful procedure, but it has two main limitations. It only has been shown to control the FDR for independent or positively correlated voxel t-statistics, and if the statistics are positively correlated, it may not be as powerful as another method which does incorporate correlation information. As discussed earlier, most fMRI datasets exhibit some spatial correlation. Therefore, it is useful to consider methods for controlling the FDR when there is correlation present. Yekutieli and Benjamini (YB, 1999) propose a parametric estimate of the FDR under correlation. Define $V(\gamma)$ to be the random variable

representing the number of inactive voxels in the true image (with some active and some inactive voxels) which are declared active using threshold γ. Similarly, define $R_0(\gamma)$ to be the random variable representing the number of inactive voxels which are declared active using threshold γ, in an image reflecting the overall null hypothesis where all voxels are inactive. Then $R_0(\gamma)$ is stochastically larger than $V(\gamma)$ because it has more inactive voxels which may potentially be declared active. Yekutieli and Benjamini (1999) use this result to construct a conservative estimator of the FDR for threshold γ as

$$E\left[\frac{R_0(\gamma)}{R_0(\gamma)+S(\gamma)}\right] \text{ instead of } E\left[\frac{V(\gamma)}{V(\gamma)+S(\gamma)}\right].$$

In this expression, we can estimate $S(\gamma)$ by

$$\hat{S}(\gamma) = R(\gamma) - mp_\gamma,$$

where $R(\gamma)$ is the number of t_i exceeding γ in the observed image and p_γ is the p-value corresponding to the threshold γ. Then the final estimator is

$$\widehat{FDR}_{YB}(\gamma) = E\left[\frac{R_0(\gamma)}{R_0(\gamma)+\hat{S}(\gamma)}\right]. \tag{4.1}$$

This expectation is evaluated using the following steps:

1. For a given threshold (or p-value) γ, estimate $\hat{S}(\gamma)$ using the observed image data.

2. Simulate a series of B images under the null hypothesis, either through nonparametric resampling or Gaussian random field as detailed earlier, and denote statistics computed on simulated image b, $b = 1, \ldots, B$ with a superscript.

3. For simulated image b compute $R_0^b(\gamma)$, the number of $|t_i^b|$'s in the bth image exceeding γ.

4. Evaluate the expectation in (4.1) using the simulated images as

$$\widehat{FDR}_{YB}(\gamma) = \sum_{b=1}^{B}\left[\frac{R_0^b(\gamma)}{R_0^b(\gamma)+\hat{S}(\gamma)}\right].$$

These FDR estimates can be used in a step-up procedure analogous to the BH procedure as follows. First order the observed p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, and let $v_{(i)}$

denote the voxel corresponding to p-value $p_{(i)}$. Compute the corresponding FDR estimates for each p-value, and denote the FDR estimate for $p_{(i)}$ as $\widehat{FDR}_{YB}(p_{(i)})$. Let r denote the largest i for which $\widehat{FDR}_{YB}(p_{(i)}) \leq \alpha$. Conclude that the voxels $v_{(1)}, \ldots, v_{(r)}$ are active, and the remaining ones are inactive.

The YB method can be used to control the FDR when there are potentially negative correlations, and it may be more powerful than the BH method because it explicitly incorporates the correlation structure.

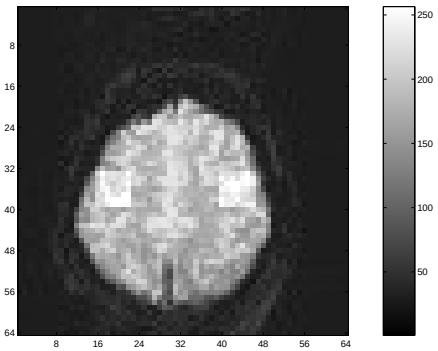
5. FMRI Simulation Study

5.1 Design

Data is generated to simulate a bilateral finger tapping fMRI block design experiment where the true motor activation structure is known so that each of the thresholding methods can be evaluated.

A 64×64 slice is selected for analysis within which two 7×7 ROI's as lightened in Figure 1 are designated to have activation. For this slice, simulated FMRI data is constructed according to a regression model which consists of an intercept, a time trend for all voxels but also a reference function for voxels in each ROI which is related to a block experimental design.

Figure 1: Anatomical with ROI.



The multivariate regression model (Rowe 2003) from which the data for all $p = 4096$ voxels and all $n = 128$ time points, is represented in terms of matrices as

$$\begin{array}{ccccccc} Y & = & X & & B & + & E \\ n \times p & & n \times (q+1) & & (q+1) \times p & & n \times p \end{array} \tag{5.1}$$

where q is the number of independent variables.

The voxels in each ROI are numbered sequentially from top left to bottom right. The simulated data is generated according to Equation 5.1 where the design matrix X is an $n \times 3$ matrix whose first column is an n dimensional vector of ones, the second column is at an n

dimensional vector of the first counting numbers, and the third column is an n dimensional vector consisting of eight replicates of eight ones then eight negative ones. The true regression coefficient matrix B outside each ROI consists of column vectors which are $(.5, .5, 0)'$ plus random independent noise for the nonzero elements with zero mean and standard deviation 0.25. Inside each ROI, the regression coefficients associated with the reference function are given in terms of (i, j) coordinates by

$$B(i,j) = 2e^{-\frac{(i-i')^2+(j-j')^2}{2(2)}} + 1.5 \tag{5.2}$$

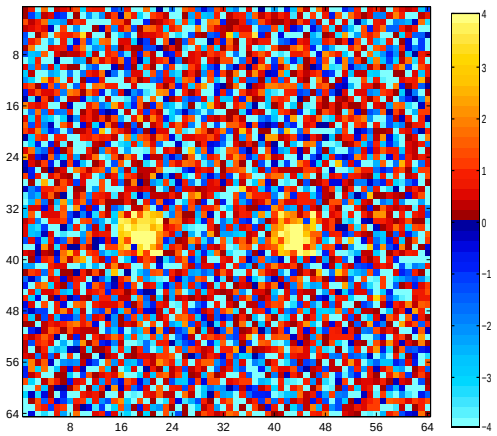
where (i', j') is the voxel number in the center of the ROI. These coefficients were chosen to have an activation region with the largest effect in the center and smaller effects towards the edge, but with reasonable power after multiplicity adjustment to detect the activations.

The voxels were assumed to have a second order stationary AR(1) spatial correlation R in which each voxel is correlated with all other voxels according to ρ^{d} where d is the Euclidean distance between the voxels. The observation errors ϵ_i were randomly generated independently from a multivariate normal distribution with p dimensional zero mean vector and $p \times p$ positive definite covariance matrix $\Sigma = \sigma^2 R$. Values for the variance and correlation were selected to be $\sigma^2 = 64$ and $\rho = 0.0$, 0.7, or 0.95.

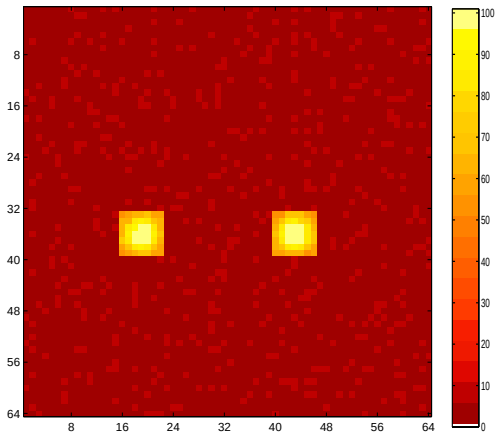
5.2 Results of Simulation Study

Five thresholding methods were considered: Unadjusted method with type I error of 5%, a Bonferroni procedure with FWE of 5%, a permutation resampling procedure to control the FWE at 5%, the BH procedure with FDR of 5%, and the YB permutation resampling procedure with a FDR of 5%. Both resampling methods used 500 randomly generated permutations. Because of the computational burden, 500 simulated images were created, on which each of these methods was applied. For each procedure, a power image was constructed which summarized the frequency over the 500 simulated images with which each voxel was detected as active (above the respective threshold). For clarity, all voxels which are never detected as active are whited out, while those which are detected active are given a color expressing the power or frequency with which they are declared active. These images are given in Figures 2, 3, and 4 for $\rho = 0.0$, $\rho = 0.7$, and $\rho = 0.95$ respectively. Also included in

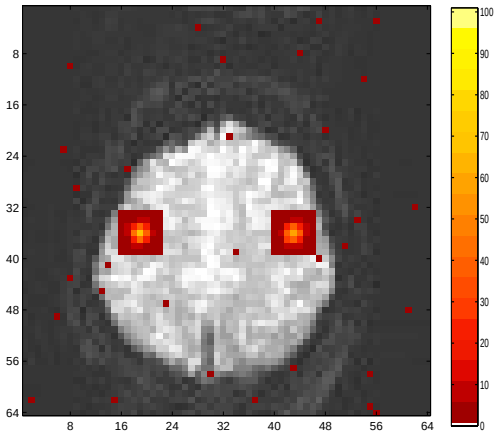
Figure 2: Sample t-statistic image and $\alpha = 0.05$ power images for $\rho = 0.00$.



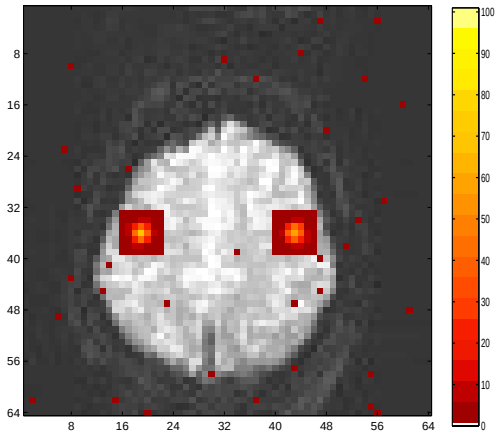
(a) Sample t-statistic image



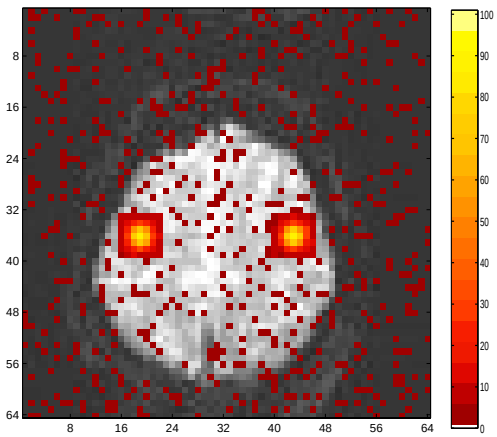
(b) Unadjusted threshold



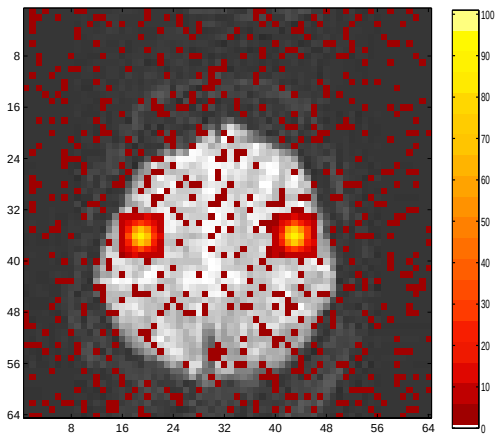
(c) FWE Bonferroni method



(d) FWE Permutation method

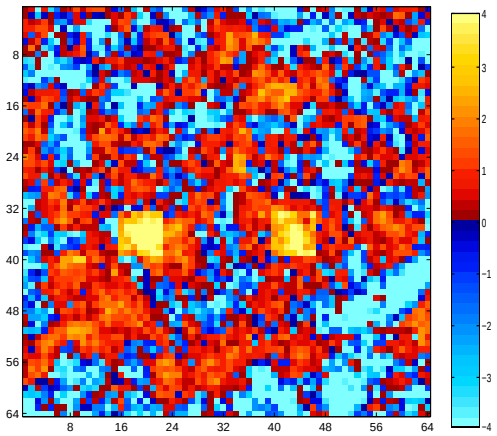


(e) FDR BH method

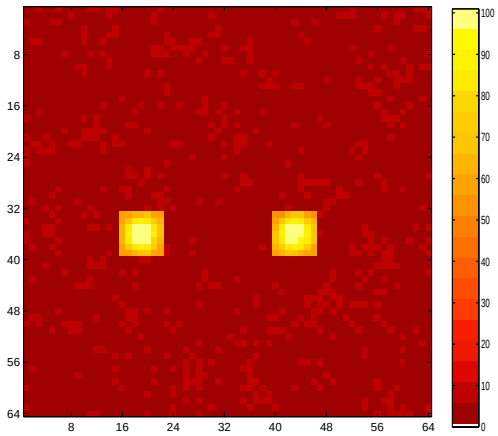


(f) FDR YB method

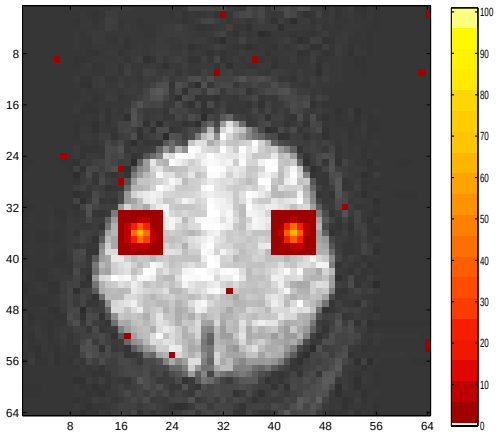
Figure 3: Sample t-statistic image and $\alpha = 0.05$ power images for $\rho = 0.70$.



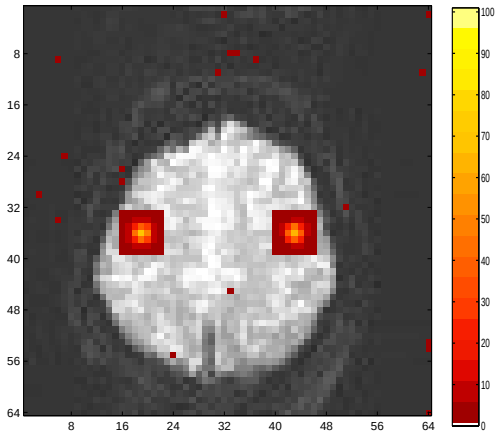
(a) Sample t-statistic image



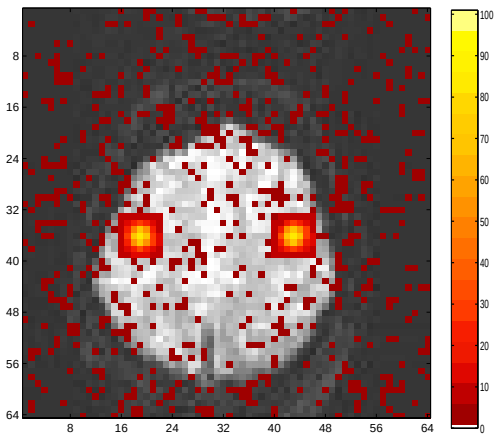
(b) Unadjusted threshold



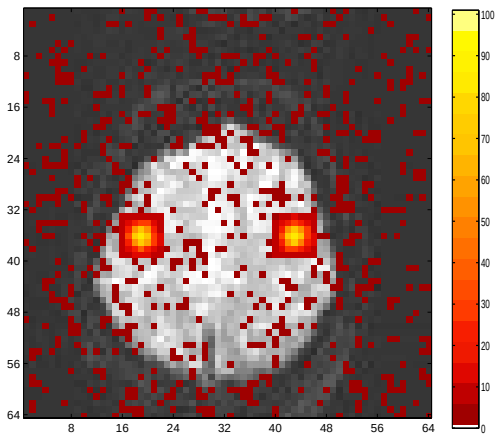
(c) FWE Bonferroni method



(d) FWE Permutation method

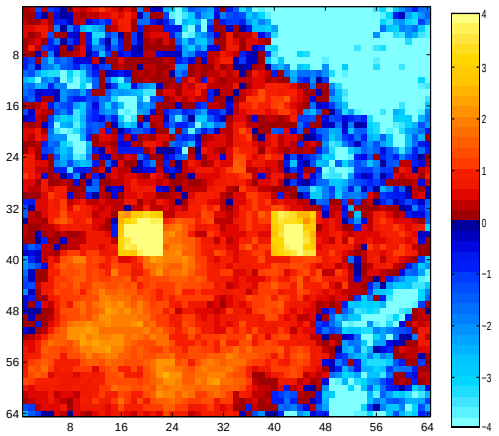


(e) FDR BH method

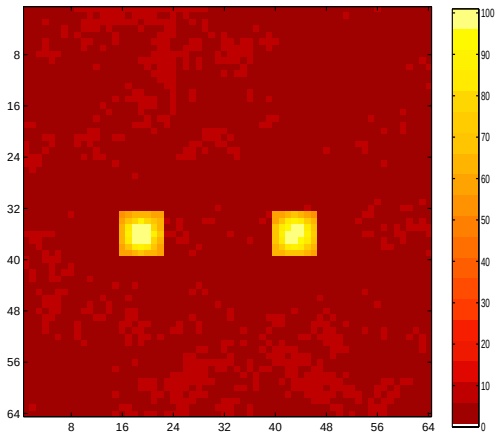


(f) FDR YB method

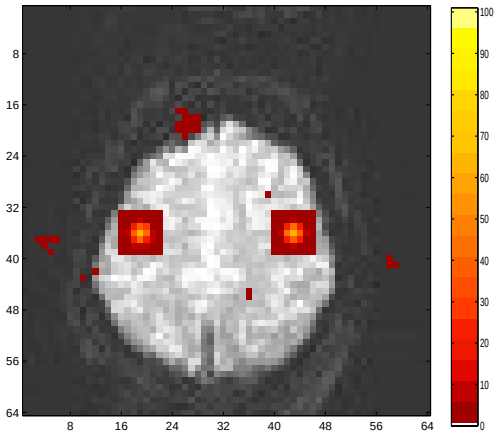
Figure 4: Sample t-statistic image and $\alpha = 0.05$ power images for $\rho = 0.95$.



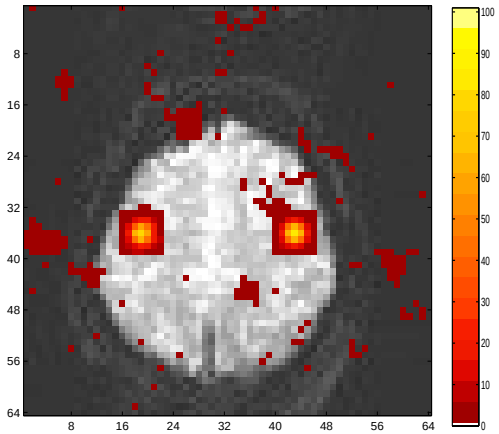
(a) Sample t-statistic image



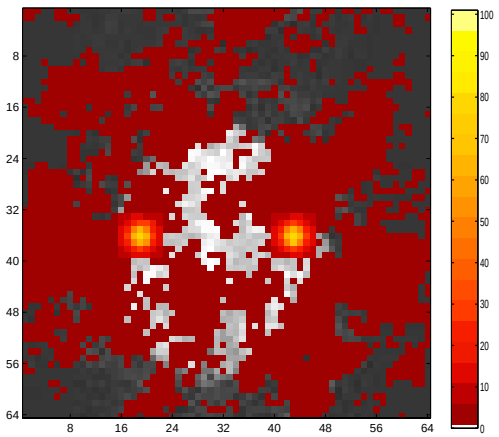
(b) Unadjusted threshold



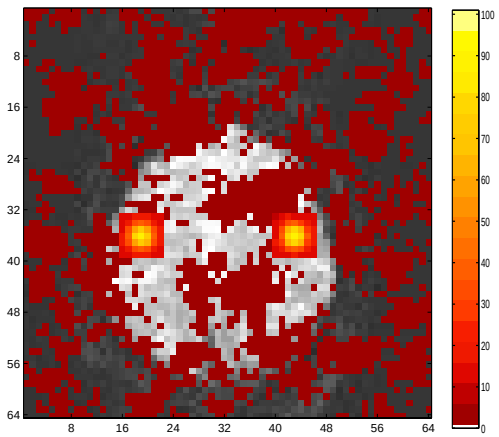
(c) FWE Bonferroni method



(d) FWE Permutation method



(e) FDR BH method



(f) FDR YB method

the first subfigure of each figure is a single sample set of observed t-statistics generated from the corresponding model and covariance matrix. These are shown to illustrate the effect of the spatial autocorrelation on how the t-statistic image appears. The image with $\rho = 0.0$ has little to no clustering of colors, the image with $\rho = 0.7$ has moderate clustering of the t-statistic values, while the image with $\rho = 0.95$ has large areas of clustering. The spatial correlation structure of most fMRI data is expected to resemble the $\rho = 0.0$ or $\rho = 0.7$ scenarios, and is not likely to be as strong as the $\rho = 0.95$ scenario.

In all three cases considered, the unadjusted method detects the active region with high power, but at the cost of a substantial number of false positives. In fact, every voxel in the entire 64 by 64 image is declared active at least once in the 500 simulated images.

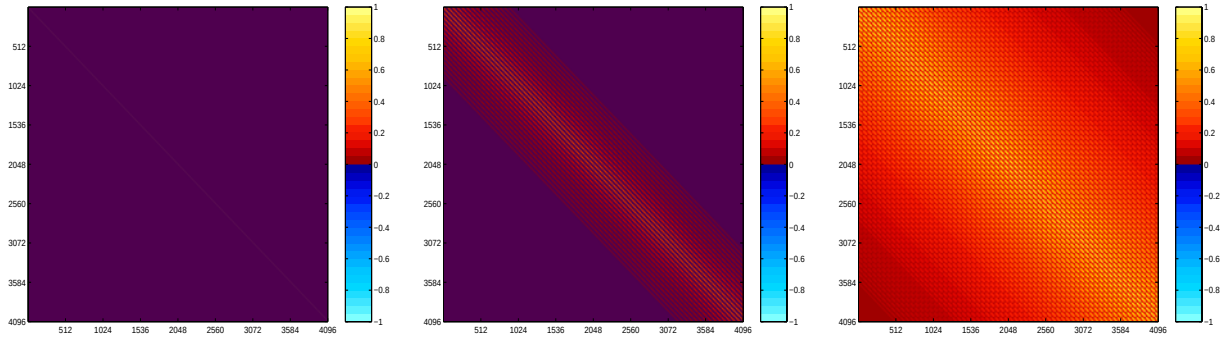
For the zero correlation scenario, there appears to be no benefit to using a permutation sampling method to account for correlation. The FWE Bonferroni method and the FWE permutation method give virtually identical power images, as do the FDR BH method and the FDR YB method. However, there is a substantial difference between a FWE-controlling procedure and a FDR-controlling procedure. The FDR-controlling procedures maintain higher power in the ROI than the FWE-controlling procedures, albeit at the cost of more falsely detected voxels. However, the proportion of falsely detected voxels is still maintained at a rate of 5% or fewer on average of the total number of voxels declared active.

Similar results can be seen for the moderately higher correlation value of $\rho = 0.7$. Even though the spatial correlation is stronger, there is still little to no effect of incorporating the correlation information through a permutation method, either for the FWE or FDR controlling procedures. This is probably because, even though $\rho = 0.7$, the correlation between any voxel and a neighbor 6 voxels away is only 0.12, which is very low. Therefore, in a 64×64 voxel image, this implies a very sparse correlation matrix in the sense that most of the values will be close to 0. Figure 5 illustrates the sparseness of the population correlation matrices for the three values of ρ by ordering the voxels from left to right and top to bottom, computing the correlation between each pair of voxels, and mapping the correlations to a color map. For $\rho = 0.7$, the correlation image has very few values which are much above 0. Since the correlation matrix is sparse, there is little advantage to incorporating such correlation information into the multiplicity adjustment. However, there is still an

important advantage in terms of power to use a FDR-controlling procedure rather than a FWE-controlling procedure, similar to when the correlation was 0.0.

When the spatial correlation is very strong, then we can see evidence that permutation resampling methods improve the power to detect voxel activations. The FWE permutation sampling procedure detects a larger portion of the activation region with higher power than the FWE Bonferroni procedure. Similarly, the FDR YB resampling method also has higher power than the FDR BH method to detect voxel activations. For the spatial correlation of 0.95, the correlation between any voxel and a neighbor 6 and 12 voxels away is 0.74 and 0.54 respectively, and this is illustrated in Figure 5 by a correlation map with a larger frequency of moderate to high correlation values. This substantially stronger spatial autocorrelation is utilized by the resampling methods to improve the power relative to their nonresampling counterparts. As above, the FDR-controlling methods are more powerful than the FWE-controlling methods. However, as indicated by the sample t-statistic image, spatial autocorrelation of this extent is unlikely to be encountered in fMRI data.

Figure 5: Population correlation images



(a) $\rho = .00$ (b) $\rho = .70$ (c) $\rho = .95$

Table 2 below provides additional information on the magnitude of the power differences between the various methods, as well as the error rates obtained.

The FDR-controlling procedures appear to improve the average power in the active region by approximately 14%. The resampling adjustments to incorporate correlation improve the power by about 7% but only when $\rho = 0.95$.

Additional correlation structures were also considered, including some with negatively

Table 2: Average Power, FDR, and FWE for various thresholding methods

	ρ		
Method	0.0	0.7	0.95
Unadjusted	0.747	0.753	0.753
FWE Bonf.	0.092	0.093	0.091
FWE Perm.	0.092	0.097	0.168
FDR BH	0.232	0.235	0.237
FDR YB	0.236	0.242	0.304

(a) power

	ρ		
0.0	0.7	0.95	
0.731	0.727	0.605	
0.007	0.004	0.008	
0.007	0.005	0.024	
0.049	0.048	0.036	
0.052	0.053	0.067	

(b) FDR

	ρ	
0.0	0.7	0.95
1.000	1.000	0.998
0.062	0.028	0.010
0.068	0.034	0.058
0.690	0.526	0.116
0.714	0.562	0.226

(c) FWE

correlated regions of the brain. In all others, similar patterns were observed, where the correlation matrix was too sparse for the correlation-based multiplicity adjustments to improve the power to detect active voxels over methods which do not account for correlation.

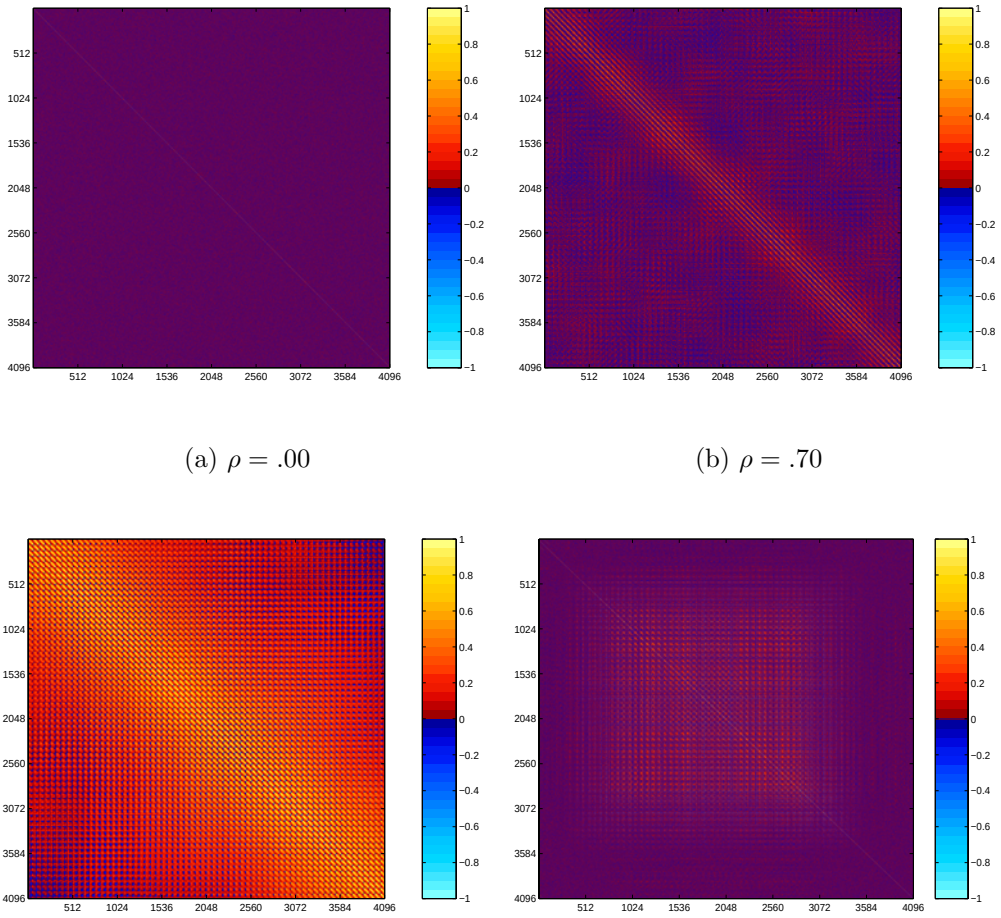
6. Real fMRI Example

To illustrate the thresholding methods described in this paper, a bilateral finger tapping experiment was performed with the same design as the previous simulation study. To generate the functional data, bilateral finger tapping was performed in a block design with eight epochs of 16s on and 16s off. Scanning was performed using a 3T Bruker Biospec in which 15 axial slices of size 64×64 were acquired. Each voxel has dimensions in mm of $3.125 \times 3.125 \times 5$, with TE= 27.2ms. Observations were taken every TR= 2000ms so that there are 128 in each voxel. Data from a single slice through the motor cortex was selected for analysis. A multiple regression model was fit to the data with an intercept, a time trend, and a reference function. The reference function was eight replicates of eight ones then eight negative ones, which mimics the experimental design in the simulation study.

After fitting the regression model, the correlation matrix was computed from the residuals and a correlation image was constructed. This sample correlation image was used to investigate the magnitude and sparseness of the spatial correlation among voxels. The image is shown in figure 6 alongside sample correlation images associated with the sample

t-statistic maps generated from the spatial autoregressive models with $\rho = 0.0, 0.7$, and 0.95, as described in the simulation study.

Figure 6: Sample correlation images



(a) $\rho = .00$

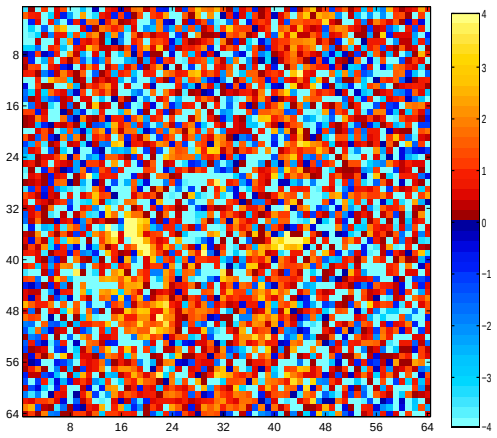
(b) $\rho = .70$

(c) $\rho = .95$

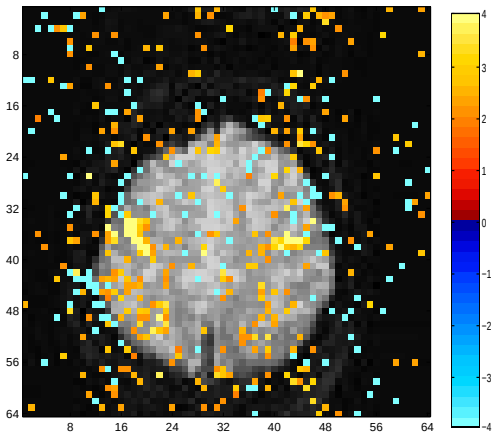
(d) real dataset

Although the sample correlation image from the real dataset does not have the same spatial correlation structure as the generated images, it appears that the magnitude and sparsity of this real dataset is similar to the situation where $\rho = 0.7$, and not as strong as when $\rho = 0.95$. Therefore, we would expect there to be little difference between using a thresholding method which does not account for spatial correlation and one which does account for spatial correlation. The real data t-statistic image is given in Figure 7, along with thresholded images using each of the methods discussed with a 5% error rate.

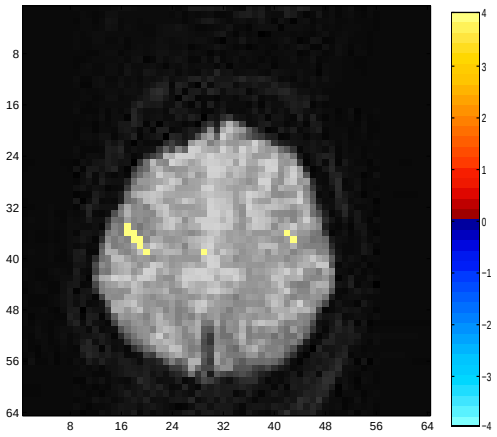
Figure 7: Real data thresholded t-statistic images for $\alpha = 0.05$.



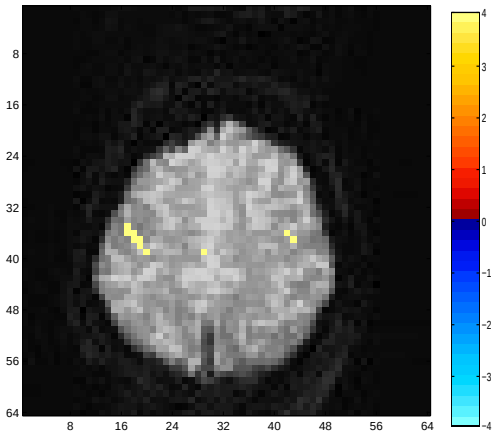
(a) Sample t-statistic image



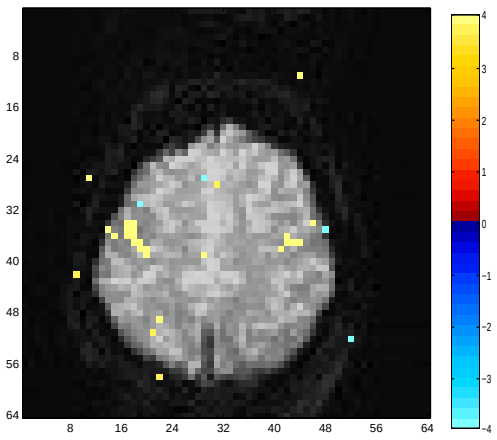
(b) Unadjusted threshold



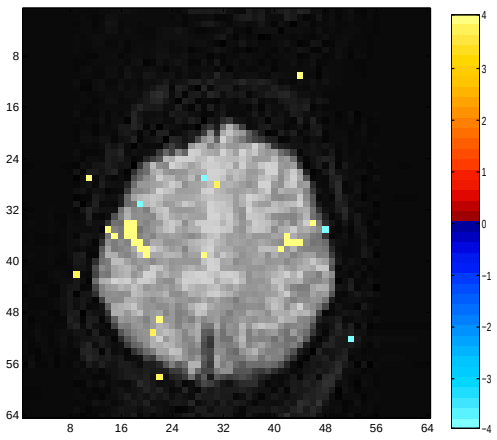
(c) FWE Bonferroni method



(d) FWE Permutation method

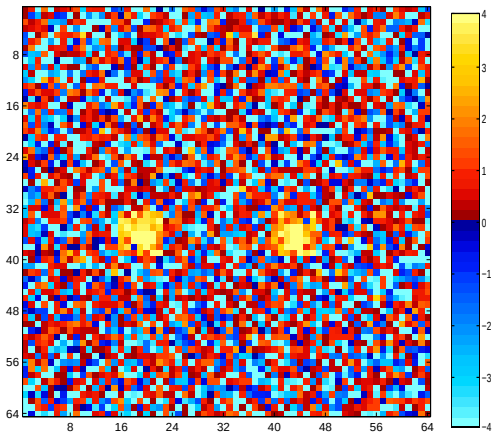


(e) FDR BH method

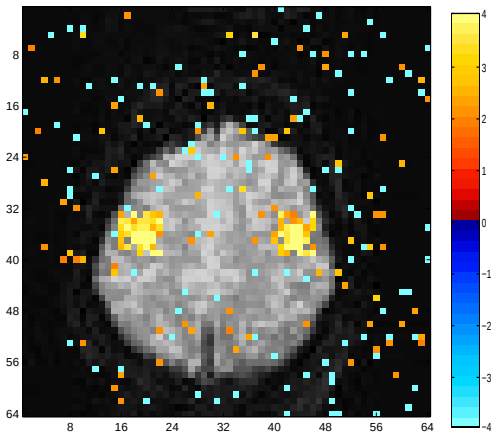


(f) FDR YB method

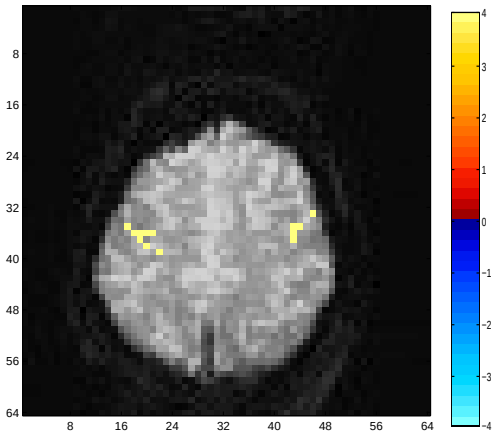
Figure 8: Sample thresholded t-statistic images for $\alpha = 0.05$ and $\rho = 0.0$.



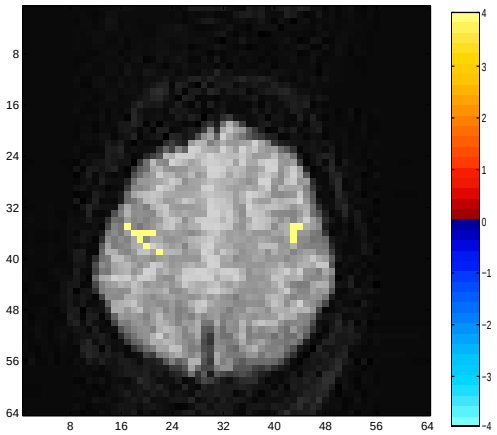
(a) Sample t-statistic image



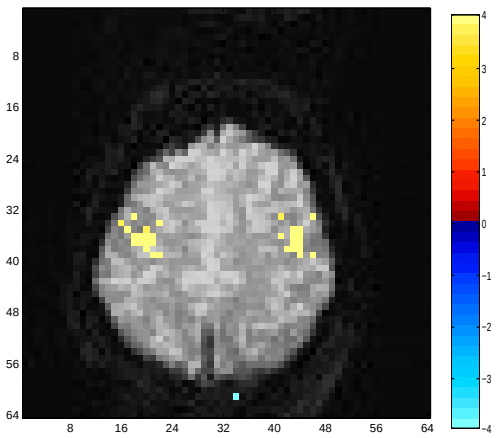
(b) Unadjusted threshold



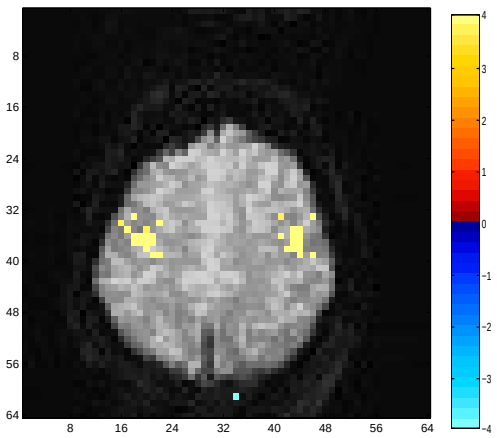
(c) FWE Bonferroni method



(d) FWE Permutation method

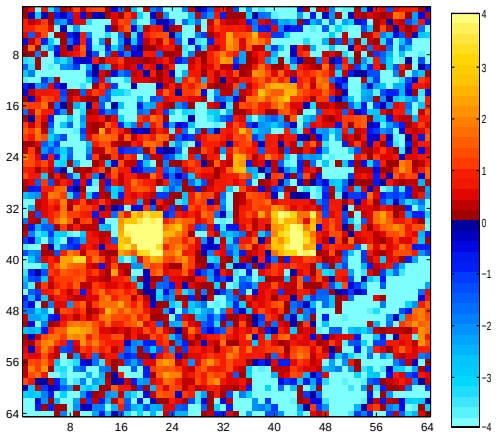


(e) FDR BH method

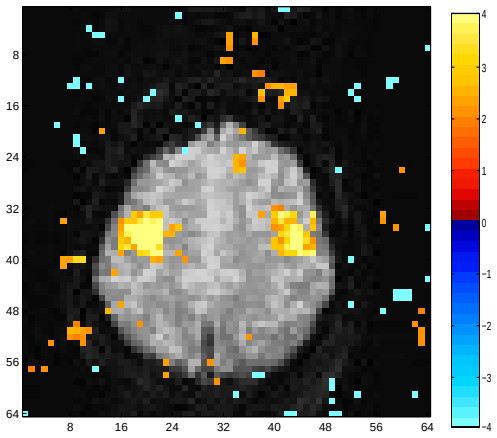


(f) FDR YB method

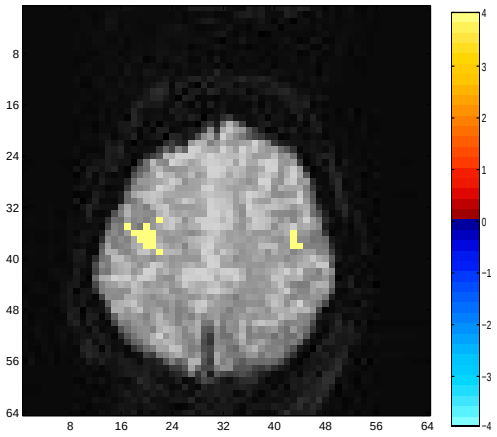
Figure 9: Sample thresholded t-statistic image for $\alpha = 0.05$ and $\rho = 0.7$.



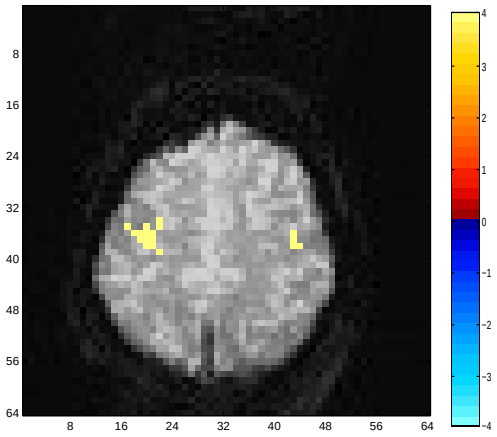
(a) Sample t-statistic image



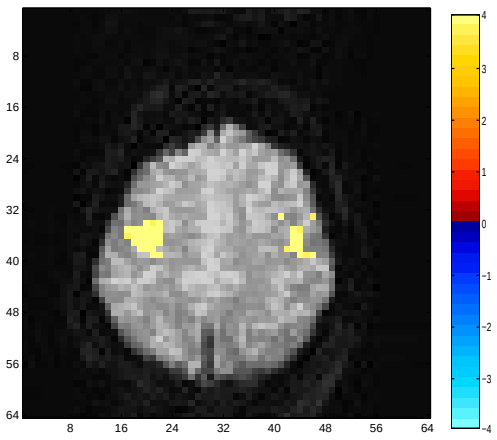
(b) Unadjusted threshold



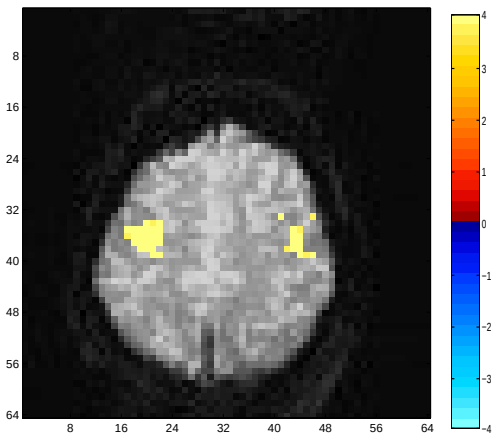
(c) FWE Bonferroni method



(d) FWE Permutation method

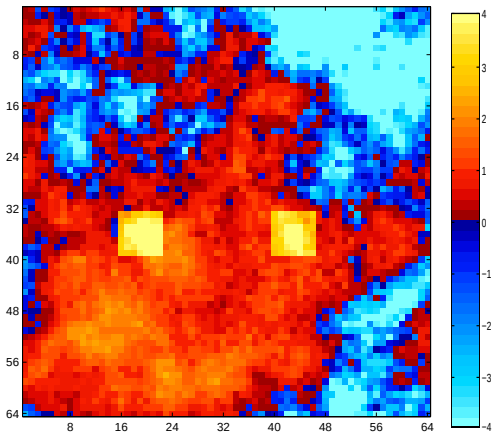


(e) FDR BH method

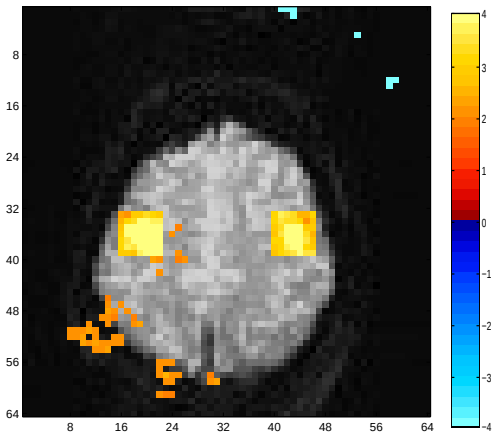


(f) FDR YB method

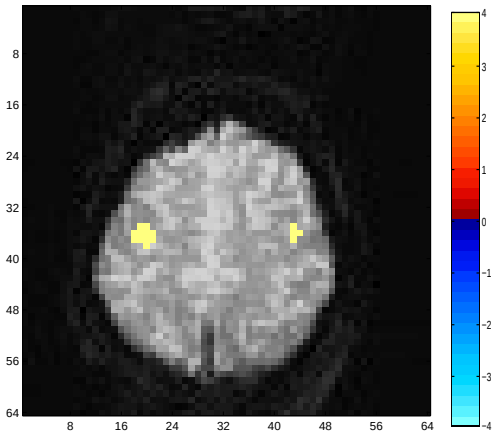
Figure 10: Sample thresholded t-statistic image for $\alpha = 0.05$ and $\rho = 0.95$.



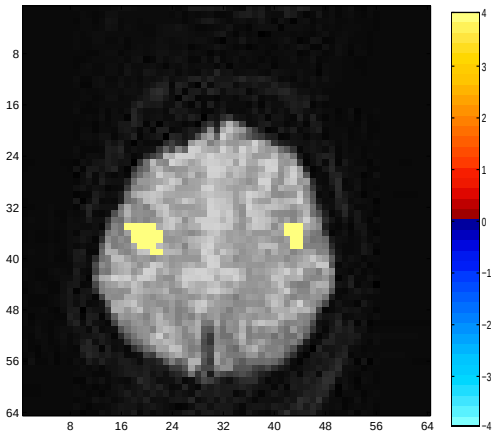
(a) Sample t-statistic image



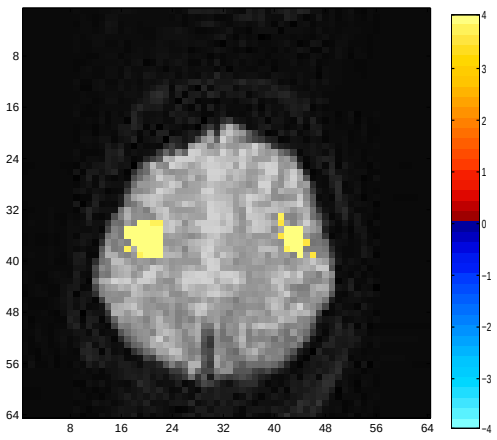
(b) Unadjusted threshold



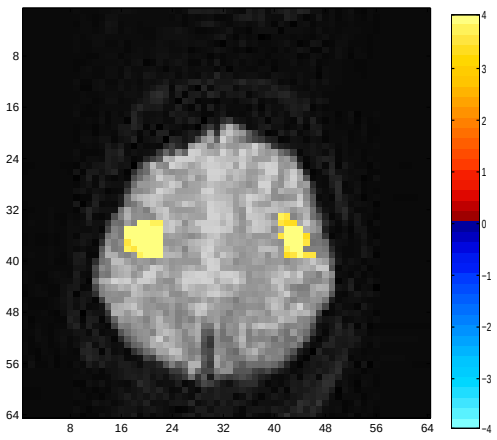
(c) FWE Bonferroni method



(d) FWE Permutation method



(e) FDR BH method



(f) FDR YB method

As expected, there are little to no differences between the Bonferroni and the permutation resampling FWE adjustment, or between the BH and the YB FDR adjustment. Also as expected, there are more voxels above the threshold using the FDR adjustment than using a FWE adjustment, because the error rate is less stringent. These results can be compared to thresholding of the $\rho = 0.0, 0.7$, and 0.95 sample t-statistic maps in Figures 8, 9, and 10. The similarity between the real and simulated thresholded results indicates that the simulations give a realistic depiction of fMRI data.

7. Conclusion

This simulation study highlights two important findings. First, as has been indicated by other authors, the FDR-controlling methods generally have higher power than FWE-controlling methods to detect active voxels. The average magnitude of this power improvement was approximately 14% in the simulations considered, but this is likely to be sensitive to the underlying parameters involved and the size of the image considered. However, the procedures are controlling two different error rates, so this higher power comes at the cost of a greater rate of false positives. For most fMRI applications, because of the large number of voxels considered, controlling the FWE is less appealing than controlling the FDR at a fixed rate α, because the number of allowable voxels which are falsely declared active is then linked to the total number of voxels declared active.

Second, except when the spatial correlation is extremely strong ($\rho = 0.95$), voxelwise thresholding methods which use resampling to account for correlation in the multiplicity adjustment do not have much impact on the power. This is probably due to the sparseness of the overall covariance matrix in most practical fMRI applications. Therefore, because of the computational burden of doing such resampling, it is probably not worthwhile to incorporate spatial correlation unless there is an indication of a high correlation between a large number of the voxels in the image.

While incorporating correlation information does not appear to be important to voxelwise thresholding rules, note that these findings do not apply to cluster thresholding methods, where an a priori cluster size is also used to set the threshold. In this case, spatial correlation information can have a significant impact on the expected cluster size, as indicated by

our sample t-statistic images in Figures 2 to 4, and the resulting t-statistic/cluster size thresholding rule needs to be sensitive to this.

Finally, our simulation study has focused on the effect of spatial correlation on the voxel-wise thresholding methods. Many fMRI datasets also include temporal autocorrelation, and while one may whiten the data as in Locascio et al. (1997) before applying a permutation resampling method, further work needs to be done to investigate whether our conclusions hold for temporally autocorrelated data as well.

References

1. Bandettini, P.A., Jesmanowicz, A., Wong, E.C., and Hyde, J.S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, **30:** 161-173.

2. Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, 289-300.

3. Benjamini, Y. and Hochberg, Y. (2000). On Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, **25:** 60-83.

4. Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics*, **29:** 1165-1188.

5. Cox, R.W., Jesmanowicz, A., and Hyde, J.S. (1995). Real-time functional magnetic resonance imaging. *Magnetic Resonance in Medicine*, **33:** 230-236.

6. Friston, K.J., Frith, C.D., Liddle, P.F., and Frackowiak, R.S.J. (1991). Comparing functional (PET) images: the assessment of significant change. *Journal of Cerebral Blood Flow and Metabolism*, **11:** 690-699.

7. Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., and Evans, A.C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, **1:** 214-220.

8. Genovese, C.R., Lazar, N.A., and Nichols, T. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *Neuroimage*, **15:** 772-786.

9. Holmes, A.P., Blair, R.C., Watson, J.D.G., and Ford, I. (1996). Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. *Journal of Cerebral Blood Flow and Metabolism*, **16:** 7-22.

10. Locascio, J.J., Jennings, P.J., Moore, C.I. and Corkin, S. (1997). Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, **5:** 168-193.

11. Petersson, K.M., Nichols, T.E., Poline, J.-B., and Holmes, A.P. (1999). Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. *Philos Trans R Soc Lond B Biol Sci*, **354:1387** 1261-1281.

12. Rowe, D.B. (2003). Multivariate Bayesian Statistics, CRC Press, Boca Raton, FL, USA.

13. Storey, J.D. (2001a). A New Approach to False Discovery Rates and Multiple Hypothesis Testing, Technical Report No. 2001-18, Department of Statistics, Stanford University.

14. Storey, J.D. (2001b). The False Discovery Rate: A Bayesian Interpretation and the q-value, Technical Report No. 2001-12, Department of Statistics, Stanford University.

15. Storey, J.D. and Tibshirani (2001). Estimating the positive False Discovery Rate Under Dependence, with Applications to DNA Microarrays, Technical Report No. 2001-28, Department of Statistics, Stanford University.

16. Worsley, K.J., Evans, A.C., Marrett, S. and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, **12:** 900-918.

17. Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, *Journal of Statistical Planning and Inference*, **82**, 171-196.