Statistics and Its Interface Volume 1 (2008) 23–32

Detecting signals in FMRI data using powerful FDR procedures

Martina Pavlicová, Thomas J. Santner and Noel Cressie

Functional magnetic resonance imaging (FMRI) has revolutionized the study of linking physical stimuli with localized brain activity. Among the challenges of working with FMRI data, they are noisy, they exhibit spatial correlation, and they are usually large containing tens of thousands of voxels of information. The notion of False Discovery Rate (FDR) has made a great impact on how to perform powerful multiple hypothesis tests to detect signals in such large multivariate data. The spatial dependence in FMRI data requires special care since, if ignored, it can lead to a loss of control of size as well as a deterioration in power of FDR procedures. This article advocates transforming the voxel-wise test statistics to wavelet space, where the coefficients are approximately uncorrelated. We demonstrate, through a series of experiments, that an FDR procedure in wavelet space enhanced by P-value adaptive thresholding (EPAT), maintains control of the size of the multiple-testing procedure and offers substantially increased power over an FDR procedure that is applied directly to the map of (spatially dependent) test statistics. The EPAT methodology, developed here for FMRI data, is generic and can be applied in other dependent data settings.

Keywords and phrases: Brain mapping, False discovery rate, Power, Sensitivity, Size, Spatial dependence, Specificity, Wavelets.

1. INTRODUCTION

Functional magnetic resonance imaging (FMRI) is a technique for creating a temporal sequence of images of the human brain. These images are based on changes in blood oxygenation, which can occur for a number of reasons, one of them being regional brain activation. In brain mapping, FMRI is used to locate regions of the brain that are activated by a specific task. In a simple brain-mapping experiment, the subject's brain is scanned rapidly, often while the subject alternates between periods of task-activation and rest. In principle, by comparing the intensities of activation and rest images, one can identify areas of the subject's brain where there is neural activity.

One common method to identify active voxels is to perform a set of voxel-wise hypothesis tests based on the difference in the measured intensities when performing the experimental task (i.e., activation) and when at rest (i.e., baseline). Traditional procedures adjust the threshold for the set of individual voxel-wise tests so that the probability of one or more Type I errors (i.e., one or more inactive voxels in the image is falsely declared to be active) is controlled to be no more than a given level α. This criterion is the so-called family-wise error rate (FWER). The simplest thresholding procedure of this sort is the Bonferroni procedure. In the FMRI setting, with test statistics from N voxels, the Bonferroni procedure declares a test statistic significant (and the corresponding voxel to be active) if the test statistic exceeds the threshold derived from the individual test having level of significance α/N (e.g., [10]). The Bonferroni procedure applied in test-statistic space is denoted by BONF, below. BONF has very low power to detect activation and does not exploit the spatial dependence among the individual test statistics. Other researchers assume various dependence models for the distribution of the voxel-wise test statistics and control the FWER. For example, [22] assume a Gaussian random field (GRF) model, which results in a procedure that is less conservative than the Bonferroni procedure. [15] introduce a nonparametric modification of such a thresholding procedure. However, there are several assumptions that need to be checked when thresholding using GRF procedures, primarily that the test-statistic image is well approximated by a GRF. Indeed, it may be necessary to smooth the test-statistic image prior to thresholding, which can change the shape of the actual activations and has the effect of reducing the spatial resolution of the resulting activation image.

In addition to modifying procedures to control the FWER, there have been proposals to modify the criterion itself. In their pioneering work, [2], denoted BH hereafter, introduced a procedure in multiple hypotheses testing that controls the false discovery rate (FDR) at a pre-specified level q. The FDR is, roughly, the expected proportion of errors among the rejected null hypotheses of no activation. More precisely, let r denote the total number of null hypotheses that are rejected by a family of tests and v denote the number of true null hypotheses that are (falsely) rejected among the N hypotheses; set $Q \equiv 0$ if $r = 0$, and $Q \equiv v/r$ if $r > 0$. The FDR is the *expected value* of Q. Given $q \in (0, 1)$ and assuming *independent* test statistics $\{T_i\}_{i=1}^N$, BH propose testing $\{H_{0i}\}_{i=1}^N$ by first ordering the

P-values, $p_{(1)} \le p_{(2)} \le \cdots \le p_{(N)}$. Then set

$$(1) \qquad L \equiv \max\left\{ i : p_{(i)} \le \frac{i}{N} q \right\},$$

if the set on the right-hand side of (1) is non-empty. Their procedure rejects none of the N hypotheses if the set on the right-hand side of (1) is empty; otherwise, it rejects those null hypotheses whose P-values are $p_{(1)} \le p_{(2)} \le \cdots \le p_{(L)}$. BH show that this method of testing the N hypotheses controls the FDR to be less than or equal to q.

Many of the recent thresholding procedures that control the FDR are either "step-down" or "step-up" procedures that carry out individual hypothesis tests sequentially according to their P-values. Step-up (step-down) procedures start from those with the largest (smallest) P-values; the BH procedure is a step-up procedure. [8] give a computationally fast procedure that controls the FDR in the FMRI setting, and [4] consider the case where the test statistics are dependent. We denote the BH procedure applied directly to the test-statistic image by BHTS, below.

The FDR criterion was also used by [16], who proposed transforming the test-statistic image to wavelet space and implementing FDR with block-wavelet thresholding; we denote the resulting procedure by BHWA, below. [17] also proposed thresholding the discrete-wavelet transform of the test-statistic image; they proposed a way to reduce the number of hypotheses tested (in wavelet space), and they obtained final images using the inverse discrete-wavelet transformation. We refer to the enhanced FDR procedure of [17] as EFDR, below. Another procedure controlling FDR in a spatial setting was suggested by [1]. They propose a two-stage hierarchical procedure that first identifies clusters of pixels where the signal might be present and then removes pixels where the signal is absent.

Several authors have considered adaptive estimation of the proportion of true null hypotheses to improve the power of the FDR-controlling procedures. [3] present an adaptive procedure where the number of true null hypotheses is estimated and used in the procedure introduced by BH. In contrast, our procedure (described in Section 2) estimates the set of true alternative hypotheses and performs an FDR-controlling adaptive procedure on these. Other papers proposing adaptive FDR procedures are those by [9], [19], and [20]. [7] introduce a weighted FDR-controlling procedure with data-driven weights. Finally, [25] developed a heteroskedastic-robust test procedure based on a resampling method, called "wild bootstrap", for assessing the statistical significance of the associations between the brain structure and covariates. Comparisons of many spatial thresholding techniques are provided by [11, 12].

We shall develop powerful procedures in this paper that base their approach on the false discovery rate (FDR) and account for the natural spatial dependence in the FMRI data. Section 2 describes the EFDR procedure in more detail, as well as a step-up modification called the P-value Adaptive Thresholding (PAT) procedure that is meant to improve the power of EFDR. This latter procedure is referred to as EPAT below. Section 3 describes a simulation experiment used to compare these two procedures, EFDR and EPAT, with the three thresholding procedures, BONF, BHTS, and BHWA. Section 4 gives details on the implementation of the various thresholding procedures. Section 5 compares the procedures through sensitivity and specificity measures, and Section 6 contains conclusions.

2. THE EFDR AND PAT PROCEDURES

The EFDR procedure is due to [17]. EFDR applies the FDR principle (in the decorrelated wavelet domain), but sharpens it by reducing the number of hypotheses being tested. This is accomplished via the following steps:

1. The image of the test statistics is represented sparsely in the wavelet domain.
2. An optimal selection of hypotheses to be tested is made in the wavelet domain using a criterion based on "generalized degrees of freedom" ([23]).

[17] conduct their tests of the null hypotheses using a *wavelet representation* of a pixellated two-dimensional spatial image. This is because the "distinctive" wavelet coefficients of a pure signal are typically clustered, both within each scale and across different scales in the wavelet domain, whereas the wavelet coefficients of white noise and spatially correlated noise are uncorrelated or approximately uncorrelated, respectively.

While [17] applied the EFDR procedure to two-dimensional (2-D) spatial fields, it is straightforward to apply it to 3-D images as in [24]. The extension does not require any theoretical modifications because 3-D wavelets are well described in the literature. However, in practice, 3-D applications will require more extensive computational resources and longer running times than those in 2-D, and hence this paper restricts attention to 2-D applications.

In the FMRI setting, consider a 2-D slice of the brain, and let $\{T_i\}_{i=1}^N$ denote the test statistics associated with the N voxels in the slice. Instead of testing the N null hypotheses $H_{01}, H_{02}, \ldots, H_{0N}$ using $T_1, T_2, \ldots, T_N$, respectively, apply a 2-D discrete wavelet transform to the test-statistic image to yield the N *observed* wavelet coefficients $\{\nu_i\}_{i=1}^N$. Let $\{w_i\}_{i=1}^N$ denote the standardized (within each scale and each orientation) observed wavelet coefficients. Then, using $\{w_i\}_{i=1}^N$ as test statistics, one desires to test

$$H_{0i} : w_i^0 = 0; \quad \text{for } i = 1, \ldots, N, \quad \text{vs} \quad H_{1i} : w_i^0 \ne 0, \\ \text{for some } i \in \{1, \ldots, N\},$$

where $\{w_i^0\}_{i=1}^N$ are the true standardized wavelet coefficients. The EFDR procedure finds an optimal subset of N^* of the original N hypotheses to be tested ($N^* \le N$), based on *generalized degrees of freedom* ([23]). The remaining $N - N^*$ hypotheses are *omitted* from future steps and their corresponding wavelet coefficients are set to 0. Finally, BH's FDR procedure, controlled at a pre-specified level q, is applied to the N^* hypotheses.

To select the optimal subset of N^* hypotheses to be thresholded, the EFDR procedure looks for information about whether $w_i^0 = 0$, based on wavelet coefficients observed at "neighboring" indices in wavelet space. The motivation for this is that pure signal is manifested in the wavelet domain by the clustering of "large" wavelet coefficients. Our implementation of the EFDR procedure used a discrete-wavelet transformation with a decomposition into three scales. The neighborhood system we used consisted of 11 neighbors, which are the nearest locations at the same scale with possibly different orientations and adjacent locations at nearby scales with the same orientation ([17]). Let w_i^* be the statistic given by the *maximum* of the absolute value of the 11 neighboring standardized wavelet coefficients at the i-th location; $i = 1, \ldots, N$. Then for N^* determined as in the following paragraph, the N^* hypotheses are selected to be those corresponding to the N^* largest of $\{w_i^*\}_{i=1}^N$. [17] prove that if BH's FDR procedure is applied at level q to these N^* hypotheses, then the EFDR procedure controls the FDR to be no greater than q.

We now show how the optimal N^* is determined. Let N_0 denote the number of true null hypotheses. When applying the EFDR procedure, the goal is to take N^* to be close to the number of true alternative hypotheses, which is $N - N_0$. A given N^* is evaluated by measuring the closeness of $\{\hat{w}_i(N^*)\}_{i=1}^N$ to the true values $\{w_i^0\}_{i=1}^N$ using the quadratic loss function,

$$(2)\quad l\Big(\{w_i^0\}, \{\hat{w}_i(N^*)\}\Big) = \frac{\sum_{i=1}^N (w_i^0 - \hat{w}_i(N^*))^2}{N} + \sigma^2,$$

where $\widehat{w}_i(N^*)$ is given by

$$(3)\quad \hat{w}_i(N^*) \equiv \begin{cases} w_i; & \text{if } H_{0i} \text{ is rejected using EFDR,} \\ 0; & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, N$, and σ^2 is the variance of the wavelet coefficients. Here $\sigma^2 = 1$, because the wavelet coefficients are standardized. [17] show that the optimal (minimum-mean-squared-error) estimator of the loss function (2) is,

$$(4)\qquad \frac{\sum_{i=1}^N (w_i - \hat{w}_i(N^*))^2 + 2g_0(N^*)\sigma^2}{N},$$

where the quantity $g_0(N^*)$ is the generalized degrees of freedom (GDF) referred to above. In practice, $g_0(N^*)$ is obtained by Monte Carlo numerical integration ([17]). Finally, straightforward numerical optimization yields an N^* that minimizes (4).

The EFDR approach depends on using BH's FDR procedure in wavelet space. Hence, in principle, it can be enhanced further by replacing BH's procedure with P-value Adaptive Thresholding (PAT), which was introduced in [14]. Consider testing the N hypotheses $H_{01}, H_{02}, \ldots, H_{0N}$ using independent test statistics $T_1, \ldots, T_N$, respectively. Let $p_{(1)} \le \cdots \le p_{(N)}$ denote the ordered P-values of the N individual tests, and let $q \in (0, 1)$ be a pre-specified level. The PAT procedure is defined as follows. Set

$$(5)\qquad N_0 \equiv \max\left\{i : p_{(i)} \le q/(N - i + 1)\right\},$$

if the right-hand set is non-empty, and set $N_0 \equiv 1$, otherwise. Further, set

$$(6)\quad K \equiv \max\left\{i : p_{(i)} \le \frac{(i - N_0 + 1)q}{N - N_0 + 1}; i = N_0, \ldots, N\right\},$$

if the set on the right-hand side is non-empty. The PAT procedure rejects none of $\{H_{0i}\}_{i=1}^N$ if the set on the right-hand side of (6) is empty; otherwise, it rejects those null hypotheses whose P-values are $p_{(1)} \le \cdots \le p_{(K)}$. Notice that if N_0 is set equal to 1, K given by (6) is equivalent to L given by (1), in which case the PAT procedure reduces to the BH procedure. [3] use a similar approach, except that (5) is replaced with a different method to determine N_0.

We expect to improve the EFDR procedure by using the PAT procedure to test the N^* null hypotheses corresponding to the N^* largest standardized observed wavelet coefficients. Hence, we have called this the EPAT procedure, and we anticipate that, without significant loss of power, the rate of false discoveries will be much smaller than that produced by the EFDR procedure. In the Appendix, we calibrate EPAT (amongst others) so that, after taking the inverse wavelet transform, the resulting test-statistic image is thresholded and its FDR is controlled.

3. COMPARISON OF THRESHOLDING PROCEDURES: DESIGN OF THE SIMULATION

This section gives the design associated with comparing five different thresholding procedures in a simulation experiment that is meant to mimic a typical 2-D slice of FMRI test-statistic images. To summarize, the procedures compared are:

- the Bonferroni procedure on the test-statistic image (denoted BONF),
- the BH procedure on the test-statistic image (denoted BHTS),
- the BH procedure in wavelet space (denoted BHWA),
- the EFDR procedure, which is in wavelet space (denoted EFDR),
- the EFDR procedure further enhanced with PAT, which is in wavelet space (denoted EPAT).

The BHWA, EFDR, and EPAT procedures account for dependence by first transforming the 2-D image of test-statistics into wavelet space, thresholding the selected wavelet coefficients, and transforming the resulting wavelet coefficients back to test-statistic-image space. The final image is a smoothing of the original test-statistic image and represents an estimate of the subject's brain activity. In contrast, BONF and BHTS yield a binary image of activation/no-activation for each voxel in test-statistic-image space. To compare the five procedures, we calibrated each of BHWA, EFDR, and EPAT so that their respective test-statistic images are thresholded in such a way that their FDRs are controlled.

The calibration study that is described in the Appendix used images containing *only* noise (i.e., no activation). For the BHWA, EFDR, and EPAT procedures, we determined the FDR level q that should be used for hypothesis testing in wavelet space, so that the FWER in test-statistic-image space is less than or equal to a pre-specified α (= .01 or .05).

3.1 Comparison criteria

Comparisons of BONF, BHTS, BHWA, EFDR, and EPAT were made on 90 simulated, artificial-activation datasets that contain both realistic activation and spatial noise. The procedures were compared with respect to the following sensitivity and specificity criteria.

Observed Voxel-wise Sensitivity: For the i-th artificial-activation dataset, let s_i denote the observed number of false null hypotheses that are rejected and $m_{1,i}$ denote the total number of false null hypotheses. Define

$$(7) \qquad \widehat{s}_{e,i} \equiv s_i/m_{1,i}; \quad i = 1, \ldots 90.$$

The quantity $\widehat{s}_{e,i}$ is the proportion of false null hypotheses in the i-th dataset that were correctly rejected; averaged over the datasets, ave$\{\widehat{s}_{e,i}\}$ is an estimate of the probability that a signal is successfully detected (i.e., is an estimate of power).

Observed Voxel-wise Specificity: For the i-th artificial-activation dataset, let u_i denote the number of non-rejected null hypotheses that are true and $m_{0,i}$ denote the total number of true null hypotheses. Define

$$(8) \qquad \widehat{s}_{p,i} \equiv u_i/m_{0,i}; \quad i = 1, \ldots, 90.$$

The quantity $\widehat{s}_{p,i}$ is the proportion of non-rejected true null hypotheses in the i-th dataset; averaged over the datasets, ave$\{\widehat{s}_{p,i}\}$ is an estimate of the probability a true null hypothesis is not rejected (i.e., is an estimate of 1−FWER).

3.2 The noise component of the simulated datasets

We compared the five procedures described in the introduction of Section 3 using artificial-activation datasets that have known strength and location of activation. To deconvolve the signal from the noise, we have to understand (and hence simulate) the noise. This subsection describes how realistic-noise datasets were generated to have the desired spatial-statistical characteristics as described by their first two moments.

We used three experimental datasets as the starting point for simulating artificial-noise datasets. These data were collected from one female and two male subjects, between 27 and 30 years of age. All images were obtained under baseline (rest) conditions, where the subjects had no experimental stimuli and they relaxed with eyes closed during scanning. The images were obtained from a 1.5-T GE Signa MRI scanner with a standard head coil.

The three "null" datasets consisted of time-sequenced brain volumes, each of length $T = 200$ time points. At each time point, a brain scan of $64 \times 64 \times 28$ voxels was obtained, where the voxel dimensions were $3.09 \times 3.09 \times 5$ mm. For additional details on how FRMI scanning is done, see [5].

Because of the heavy computational requirements, only one two-dimensional slice (slice 15, counting from the bottom of the brain) out of 28 in the volume, was considered in the simulations described below. To simulate the noise component at each voxel, the mean, the standard deviation, and the temporal correlations at lags 1 through 14 were obtained from the selected slice, for each of the three datasets. Then 30 different artificial-noise datasets were generated from each subject's data by simulating a Gaussian time series with that subject's mean, standard deviation, and temporal correlations. This resulted in a total of 90 artificial-noise datasets, to be used in the generation of the 90 artificial-activation datasets below. (One subject's 30 artificial-noise datasets were also used in the calibration study in the Appendix.)

3.3 The activation component of the simulated datasets

To assess the effectiveness of the implemented procedures, datasets with known location and known strength of true activation were required. These were constructed with the noise component described in Section 3.2 and a visual signal component that was calibrated against images acquired from unrelated FMRI activation experiments and scaled to be between 0 (no activation) and 1 (strongest activation); see Figure 1(a).

The spatial signal alternates over the $T = 200$ time points, with 10 consecutive time points of activation followed by 10 consecutive time points at rest. The average peak signal change, defined to be the ratio of the average of the intensities under activation to the average of the intensities during the rest periods for the most activated voxel, was specified to be 3% of the expected noise.

In what follows, we make the realistic assumption that the artificial-activation datasets have non-additive noise. At each voxel, data were generated according to the formula,

$$\pi \times A \times P[t] \times E(N) + \rho \times A \times P[t] \times (N[t] - E[N]) + N[t],$$

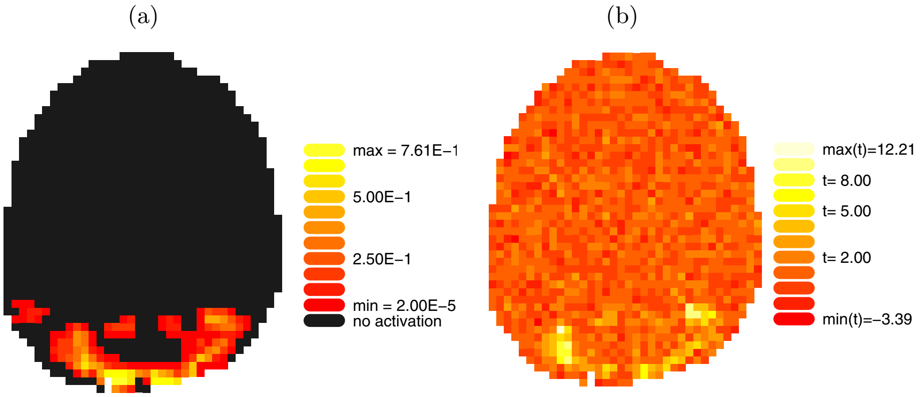

Figure 1. (a): True Signal to Be Incorporated with Each Artificial-Noise Dataset. (Yellow: Highest Activation; Red: Very Low Activation; Black: No Activation.) (b): Test-Statistic Image Corresponding to the Signal in (a), for One Artificial-Activation Dataset After Pre-Processing and Using a GLM. (Yellow: Test Statistic with High Values; Red: Test Statistic with Low Values.)

where the artificial noise $N[t]$ is the only source of randomness; $t = 1, \ldots, 200$. Here, $E(N)$ is an estimate of the expected value of the noise over the 200 time points, A is the 'activation' value between 0 and 1 shown in Figure 1(a), $P[t]$ is a value between 0 and 1 based on the experimental paradigm at time t, π is the average peak signal change, and ρ is the rate of the non-additive part of the activation component. We selected $P[t]$ to be the convolution of a box-car function (alternating between 0 and 1 every 10 time points) with a gamma function to represent the haemodynamic response (see [21]), and we set $\pi = .03$ and $\rho = .05$.

4. IMPLEMENTATION OF THE THRESHOLDING PROCEDURES

Each of the 90 artificial-activation datasets was preprocessed using FEAT (part of the FSL software package; see [18]) and then analyzed using a general linear model (GLM). In contrast to cluster-based thresholding, the data were not spatially smoothed during pre-processing. As a result, 90 test-statistic images (where Student's t-statistic was used, such as in [13]) were obtained; for example, Figure 1(b) displays one such image.

Each test-statistic image was directly thresholded using the BONF and BHTS procedures, resulting in a binary map. Figure 4(a) and Figure 4(b) show the results of applying BONF and BHTS, respectively, to the test-statistic image shown in Figure 1(b).

The BHWA, EFDR, and EPAT procedures transform the test-statistic image into wavelet space. There is a technical problem that occurs when transforming to wavelet space from test-statistic-image space (e.g., the image shown in Figure 1(b)): The parts of the image outside the brain, which have very different values than brain boundary voxels, cause the wavelet coefficients corresponding to the edge of the brain to be large. Our goal is to prepare the test-statistic image in such a way that any observed wavelet coefficient having a large value, corresponds to an observed signal in the brain rather than to a voxel's exhibiting edge effects. To solve this edge-effect problem, we blended the image of the brain into a simulated background that de-emphasizes the edge of the brain. Conditional simulation was used to extend the brain image, as illustrated in Figure 2. A conditional simulation of the brain image is a simulation from a process on the rectangle that is conditioned to be equal to the data in the original brain image, and has the same mean and covariance as the original process producing the data. Technically, it can be obtained as the sum of the simple kriging predictor based on the brain data and the residual (difference) between the (unconditionally) simulated process having the same mean as the brain-voxel data and the kriging predictor based on the simulated brain-voxel data (see Equations (3.6.20) and (3.6.21) of [6]). Conditioning on the values of the voxels that belong to the brain in Figure 2(a), leaves these voxels' values unchanged in Figure 2(b), while values of voxels from areas outside the brain have the same statistical properties (mean, variance and spatial covariance) as the in-brain voxels. The final test-statistic image is shown in Figure 2(b).

The expanded test-statistic images are transformed into wavelet space using the discrete wavelet transformation. The wavelet image is then thresholded using the BHWA, EFDR, and EPAT procedures, and the thresholded image is transformed back into test-statistic-image space using the inverse discrete-wavelet transformation. Figure 3 shows the smoothed test-statistic image (estimated signal) corresponding to the BHWA, EFDR, and EPAT procedures. There remains a final thresholding of the smoothed test-statistic image to produce a binary image, and it is based on the calibration study described in the Appendix.

Figure 4 shows the final binary (activation/no-activation) results for the five thresholding procedures (BONF, BHTS, BHWA, EFDR, EPAT) applied to the test-statistic image of Figure 1(b). The true-activation voxels are also given in Figure 4(f) and enable a visual comparison of the five procedures. A cursory comparison shows that the BONF and

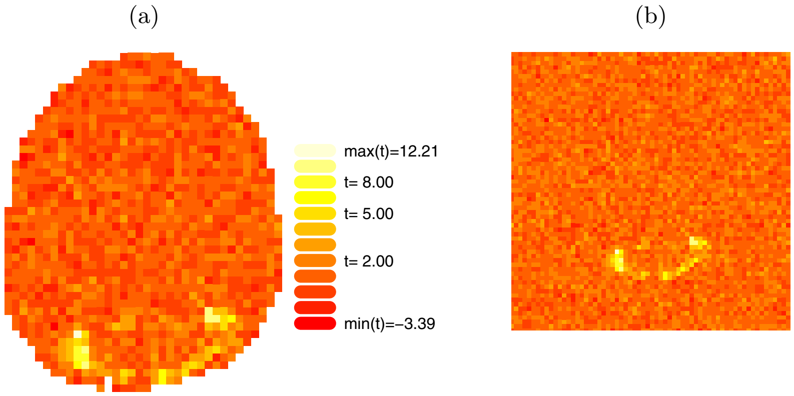

Figure 2. (a) Test-Statistic Image Shown in Figure 1(a). (b) Test-Statistic Image After Conditional Simulation, Conditioning on the Image in (a). (Yellow: Test-Statistic with High Values; Red: Test Statistic with Low Values.)

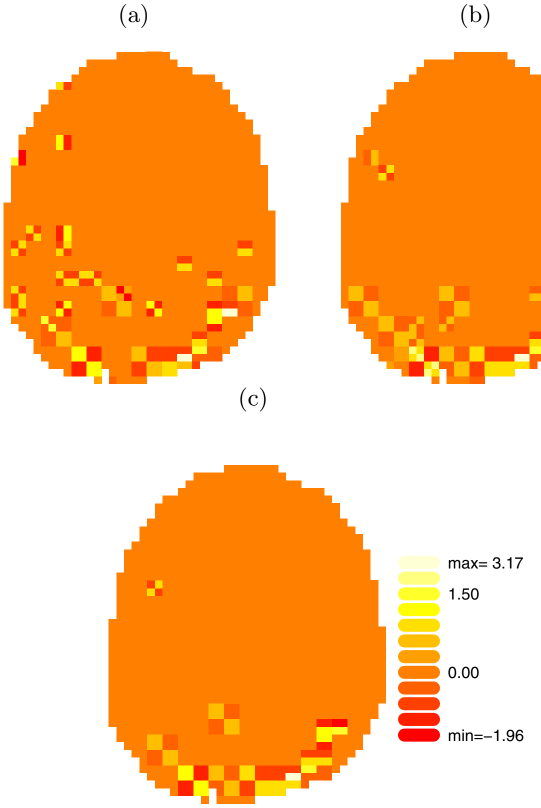

Figure 3. Images in the Test-Statistic-Image Space After Applying the Inverse Discrete Wavelet Transform to the Thresholded Wavelet Images. Shown Are Smoothed Test-Statistic Images Obtained from (a) the BHWA Procedure, (b) the EFDR Procedure, and (c) the EPAT Procedure.

BHTS procedures, neither of which specifically takes into account spatial dependence, are extremely conservative and declare only a few voxels to be activated. The next section makes a more formal comparison of the five procedures using their estimated powers and estimated FWERs based on the 90 artificial-activation datasets.

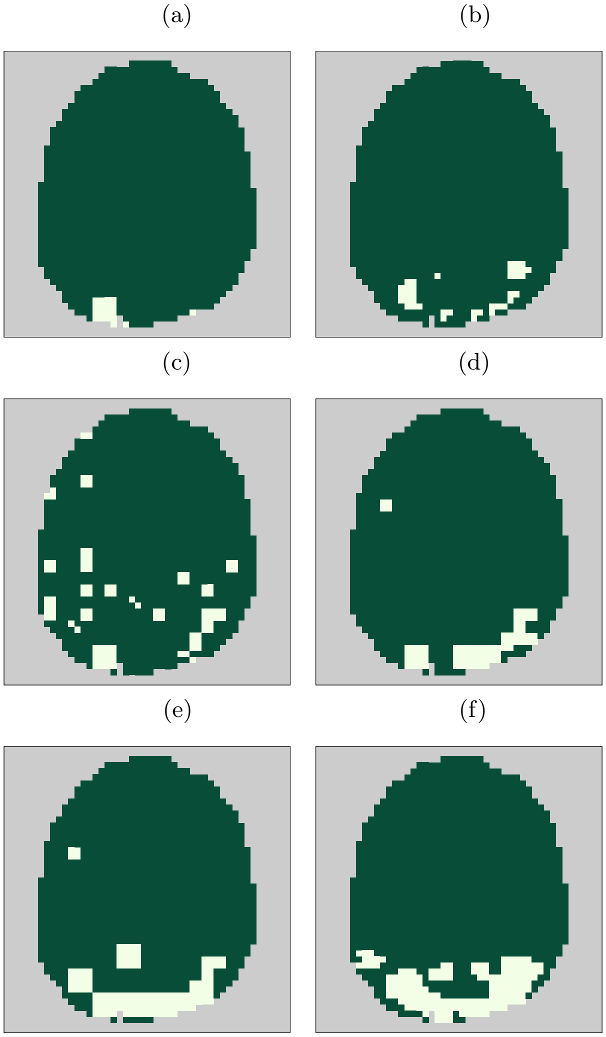

Figure 4. Binary Images Resulting from Five Thresholding Procedures: (a) BONF; (b) BHTS; (c) BHWA; (d) EFDR; and (e) EPAT. The True Activation Voxels Are Shown in (f).

5. SENSITIVITY AND SPECIFICITY COMPARISONS

Our goal is to find that procedure among BONF, BHTS, BHWA, EFDR, and EPAT having the largest power, namely $1 - Pr$ (Type II error), while having FWER, namely Pr (Type I error), below a given nominal level of significance α ($\alpha = 0.01$ or 0.05, in what is to follow). Figure 5 shows the distributions of the observed sensitivities (or power) $\{\widehat{s}_{e,i}\}_{i=1}^{90}$ and the observed specificities (or 1−FWER) $\{\widehat{s}_{p,i}\}_{i=1}^{90}$, for $\alpha = 0.01$ and 0.05, obtained from the 90 artificial-activation datasets. Each panel provides comparative boxplots of these values for all five procedures. The BONF and BHTS procedures, which are applied directly to the test-statistic images, are separated by a vertical line from the BHWA, EFDR, and EPAT procedures based on the discrete wavelet transformation. For the BHWA, EFDR, and EPAT procedures, q is the FDR level when thresholding the hypotheses in wavelet space, which is in contrast to the significance level α used in test-statistic-image space.

The distribution of the observed specificities are given in Figure 5(a) and 5(b). Because the *observed specificity* gives the proportion of true null hypotheses that are declared not significant, this value estimates $1 - Pr$ (Type I error), or 1−FWER, and hence it should be large, although procedures with observed specificities very close to 1 can result in overly conservative thresholded activation images (e.g., BONF). Small observed specificity means that the procedure is detecting only a small proportion of the true null hypotheses as not significant. Figures 5(a) and 5(b) show that both BONF and BHTS can be very conservative, and that BONF is the more conservative of the two. Among the three wavelet-based procedures, BHWA has the smallest average observed specificity of around 0.90, while EPAT has the largest average observed specificity of around 0.98, for both nominal levels of significance $\alpha = 0.01$ and 0.05. The question remains, does EPAT achieve its high specificity at the expense of giving a conservative estimate of the activation voxels? We see below that this is *not* the case.

The distribution of the observed sensitivities are given in Figures 5(c) and 5(d). Because the *observed sensitivity* is the proportion of false null hypotheses that are declared significant, this values estimates $1 - Pr$ (Type II error), or power, and hence it is desired to be large. Figures 5(c) and 5(d) show that BONF and BHTS have the smallest observed sensitivities (i.e., smallest power) among the five procedures, as one would expect, given their very large observed specificities. Both procedures detect only very small proportions of false null hypotheses as active. For the nominal level of significance $\alpha = .01$, BHWA displays the largest average sensitivity and smallest spread among the distributions of sensitivities for the three wavelet-based procedures (Figure 5(c)), although only slightly larger than EFDR and EPAT. For the nominal level of significance $\alpha = .05$, EFDR displays the

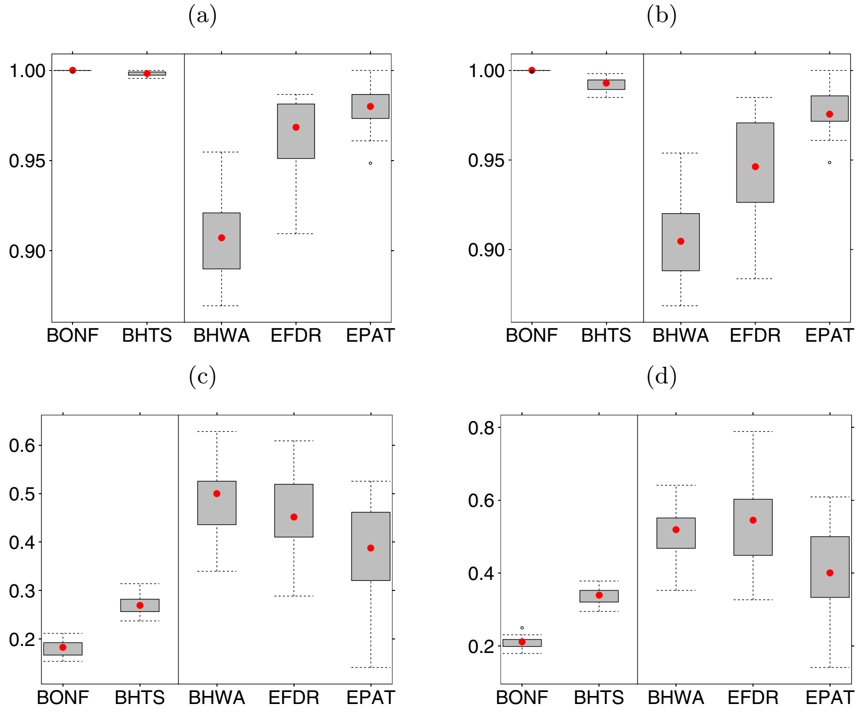

Figure 5. (a) Distribution of Observed Specificities (1−FWER) $\{\widehat{s}_{p,i}\}_{i=1}^{90}$, for $\alpha = .01$. (b) Same as in (a), Except that $\alpha = .05$. (c) Distribution of Observed Sensitivities (Power) $\{\widehat{s}_{e,i}\}_{i=1}^{90}$, for $\alpha = .01$. (d) Same as in (c), Except that $\alpha = .05$. Shown Are Box Plots with the Sample Mean Given by a Red Dot.

Table 1. Euclidean Distance of the Centroid of Each Procedure's $\{(\widehat{s}_{p,i}, \widehat{s}_{e,i})\}$-Values to the Optimal (1,1) Point. The Bold Value in Each Column Is the Smallest Distance for a Given Level of Significance α

	Distance to $(1,1)$	
Procedure	$\alpha = 0.01$	$\alpha = 0.05$
BONF	0.8186	0.7975
BHTS	0.7291	0.6641
BHWA	**0.5197**	0.4982
EFDR	0.5471	**0.4644**
EPAT	0.6358	0.5981

largest average observed sensitivity (Figure 5(d)), although only slightly larger than those of BHWA and EPAT; among the three, BHWA has the smallest spread in observed sensitivities.

Because we desire that the specificity and sensitivity be simultaneously large, an additional method of assessing the quality of each procedure is by means of the distances of (specificity, sensitivity) pairs from the desired optimal value of $(1,1)$ for the 90 artificial-activation datasets. Table 1 lists the distances from $(1,1)$ to the centroid of these pairs, for each procedure and each $\alpha \in \{.01, .05\}$. As anticipated, the procedures BONF and BHTS, which both have the largest marginal specificities and the smallest marginal sensitivities, have the greatest distances to $(1,1)$.

The BHWA and EFDR procedures have non-dominated distance measures; BHWA is closer to $(1,1)$ when $\alpha = 0.01$, while EFDR is closer to $(1,1)$ when $\alpha = 0.05$. Both BHWA and EFDR are closer to $(1,1)$ than EPAT for both α-levels. If one desires to use a procedure whose specificity is controlled, that is above a given level, then EPAT is the best choice. Figure 5 makes clear that the price of specificity vigilance is slightly lower sensitivity values for some images.

6. CONCLUSIONS

All three wavelet-based procedures (BHWA, EFDR, EPAT) have larger power than the procedures applied directly to the test-statistic image (BONF, BHTS). Among the wavelet-based procedures, BFWA and EFDR appear to have larger power than EPAT. However, when carrying out hypothesis testing, the FWER is controlled to be less than or equal to a nominal level of significance α, and the best procedure is chosen as the one with the largest power from all those that satisfy this FWER control. Based on Figure 5, BONF, BHTS, and EPAT remain under consideration after the FWER control is applied. Then, from among these three remaining procedures, EPAT is easily the most powerful.

ACKNOWLEDGMENT

This research was supported by an Office of Naval Research grant, N00014-05-1-0133, and by the National Science Foundation under grant DMS-0706997, grant DMS-0406026, and Agreement No. 0112050. The authors would like to thank Antonio Algaze and Petra Schmalbrock for providing the FMRI data and the members of FMRIB, Oxford UK, for initial consultation about simulating activation FMRI datasets. We would also like to thank the referees and Editor for helpful comments that led to improvements in the presentation and content of this paper.

APPENDIX: CALIBRATION STUDY

The purpose of the calibration study was to obtain a set of thresholding parameters that transformed the smoothed test-statistic images that result from using BHWA, EFDR, and EPAT, into binary images where each voxel is declared activated or not. The goal is to choose the thresholding parameters q and κ, which are described below, so that when the procedures are used on images that contain only noise, the average false-positive rate (an estimate of the FWER) is less than or equal to a given level of significance α. In this calibration study, the false-positive rate for an artificial-noise image is the proportion of voxels that were declared active; clearly, it can be interpreted as an empirical FDR. The calibration was carried out on a set of 30 images that contained only noise (Section 3.2 describes the method of construction of such images), which we call artificial-noise datasets.

In more detail, the thresholding parameters are

- q: the FDR level in wavelet space that controls the BH, respectively PAT, thresholding inside the EFDR, respectively EPAT, procedure. We consider the cases $q = 0.01, 0.05, 0.10, 0.20, \ldots, 0.90, 0.95, 0.99$.
- κ: the threshold applied in test-statistic-image space to the inverse-discrete-wavelet-transformed image.

For a given procedure and a given q, κ was selected so that the average false-positive rate for the 30 artificial-noise datasets was just below the desired level of significance α. The quantity κ was determined numerically by starting with a sufficiently large value so that none of the voxels of the 30 artificial-noise datasets was significant (and thus the average false-positive rate was equal to zero). Then κ was decreased in units of 0.1 and the average false-positive rate from the 30 artificial-noise datasets was recomputed, giving a non-decreasing false-positive rate. This process was continued until κ produced a false-positive rate just above α, resulting in a value of κ for each q (and α). In Figure 6, we illustrate the calibration methodology using the EPAT procedure. The top graph of Figure 6(a) plots the computed κ versus q, for $\alpha = 0.01$, and the top graph of Figure 6(b) plots the computed κ versus q, for $\alpha = 0.05$. The bottom graph in each panel shows the distribution of the proportions of true null hypotheses that were found to be falsely significant (i.e., the false-positive rate) over the 30 artificial-noise datasets. The horizontal dotted line in each plot is at height α, and the square dot denotes the average of the 30 false-positive rates.

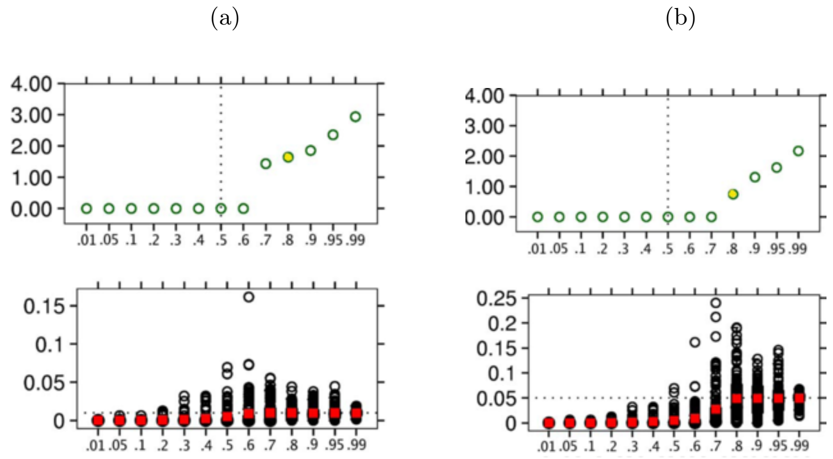

Figure 6. Results from the Calibration Study, Shown Here for the EPAT Procedure. (a) The Nominal Level of Significance Is $\alpha = .01$; the Top Plot Shows κ on the Vertical Axis Versus q on the Horizontal Axis, and the Bottom Plot Shows the Empirical FDR for the 30 Artificial-Noise Images Versus q on the Horizontal Axis. In the Bottom Plot, the Dotted Line Is at Height α and the Red Squares Are Averages Over the 30 Images. (b) The Nominal Level of Significance Is $\alpha = .05$; Plots Are the Same as in (a).

Table 2. Selected Values of κ Chosen to Yield, on Average, a False-Positive Rate Less than or Equal to α for the BHWA, EFDR, and EPAT Procedures. Here, $q = .7$ for All Three Procedures

α	BHWA	EFDR	EPAT
0.01	1.566	1.357	0.571
0.05	1.090	0.889	0

For example, one can see that when $\alpha = .01$ and $q = 0.08$, the value of κ was determined to be about 1.8. One can also visually evaluate the spread of the false-positive rates in this case as ranging from about 0.01 to 0.05.

The final choice of (q, κ) was based on the spread *and* the average of the observed proportions of true null hypotheses that were declared significant. For all procedures, we used $q = .7$, and we selected the procedure-specific κ according to the values listed in Table 2.

Received 8 January 2008

REFERENCES

[1] Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association* **102** 1272–1281.

[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, Series B* **57** 289–300. MR1325392

[3] Benjamini, Y. and Hochberg, Y. (2000). On the 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. MR1869245

[5] Buxton, R. B. (2002). *Introduction to Functional Magnetic Resonance Imaging*. Cambridge University Press, Cambridge, UK.

[6] Cressie, N. (1993). *Statistics for Spatial Data*, revised edition. Wiley, New York. MR1239641

[7] Finos, L. and Salmaso, L. (2007). FDR- and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference* **137** 3859–3870. MR2368532

[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** 870–878.

[9] Genovese, C. R. and Wasserman, L. (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society, Series B* **64** 499–517. MR1924303

[10] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York. MR0914493

[11] Logan, B., Geliazkova, M., and Rowe, D. (2008). An evaluation of spatial thresholding techniques in fMRI analysis. *Human Brain Mapping*, in press.

[12] Logan, B. and Rowe, D. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage* **22** 95–108.

[13] Pavlicová, M., Cressie, N., and Santner, T. J. (2006). Testing for activation in data from FMRI experiments. *Journal of Data Science* **4** 275–289.

[14] Pavlicová, M., Cressie, N., Santner, T. J., and Algaze, A. (2003). Using enhanced FDR to find activation in FMRI images. *NeuroImage* **19** S919.

[15] Poline, J.-B., Worsley, K. J., Evans, A. C., and Friston, K. J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage* **5** 83–96.

[16] Raz, J. and Turecky, B. (1999). Wavelet ANOVA and fMRI. In *Wavelet Applications in Signal and Image Processing VII, Proceedings of SPIE* **3813** 561–570. SPIE, Bellingham, WA.

[17] Shen, X., Huang, H.-C., and Cressie, N. (2002). Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association* **97** 1122–1140. MR1951265

[18] Smith, S. M., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, M., Leibovici, D., Ripley, B., Woolrich, M., and Zhang, Y. (2001). FSL: New

tools for functional and structural brain image analysis. *NeuroImage* **13** S249.

[19] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64** 479–498. MR1924302

[20] Storey, J. D., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* **66** 187–205. MR2035766

[21] Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith, S. M. (2004). Fully Bayesian spatio-temporal modeling of FMRI data. *IEEE Transactions on Medical Imaging* **23** 213–231.

[22] Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* **4** 58–73.

[23] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93** 120–131. MR1614596

[24] Zeng, L., Jansen, C., Unser, M. A., and Hunziker, P. (2001). Extension of wavelet compression algorithms to 3D and 4D image data: exploitation of data coherence in higher dimensions allows very high compression ratios. In *Wavelets: Applications in Signal and Image Processing IX, Proceedings of SPIE* **4478** 427–433. SPIE, Bellingham, WA.

[25] Zhu, H., Ibrahim, J., Tang, N., Rowe, D., Hau, X., Bansal, R., and Peterson, B. (2007). Statistical analysis of brain morphometric measures: A wild bootstrap method. *IEEE Transactions on Medical Imaging* **26**(7) 954–966.

Martina Pavlicová
Department of Biostatistics
Columbia University
New York, NY 10032, USA
E-mail address: mp2370@columbia.edu

Thomas J. Santner
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA
E-mail address: santner.1@osu.edu

Noel Cressie
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA
E-mail address: ncressie@stat.osu.edu